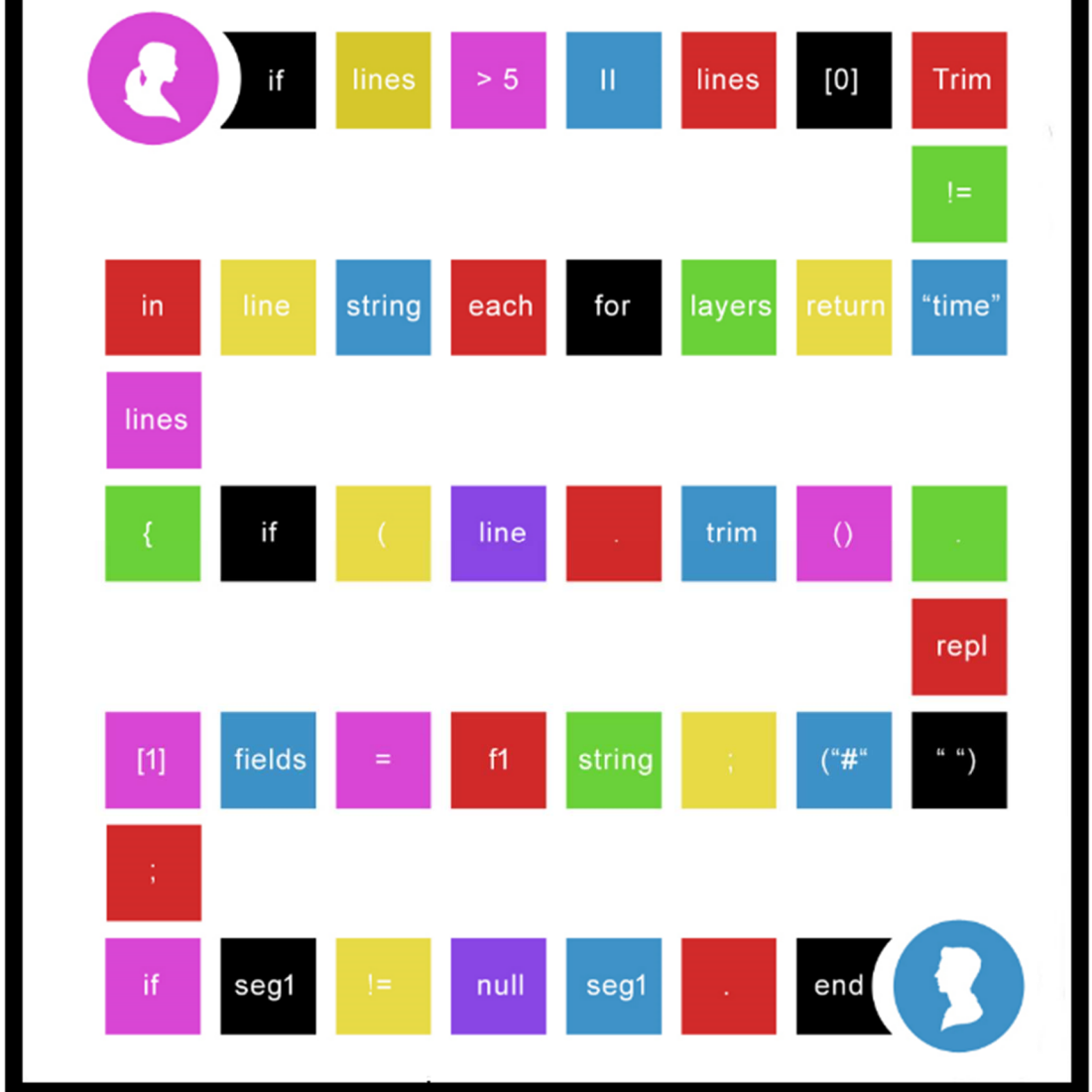
the Phonetician

A Publication of ISPhS/International Society of Phonetic Sciences



NUMBER 111-112 2015/I-II

ISPhS
International Society of Phonetic Sciences

President: Ruth Huntley Bahr

Secretary General:
Mária Gósy

Honorary President:
Harry Hollien

Vice Presidents:
Angelika Braun
Marie Dohalská-Zichová
Mária Gósy
Damir Horga
Heinrich Kelz
Stephen Lambacher
Asher Laufer
Judith Rosenhouse

Past Presidents:
Jens-Peter Köster
Harry Hollien
William A. Sakow †
Martin Kloster-Jensen†
Milan Romportl †
Bertil Malmberg †
Eberhard Zwirner †
Daniel Jones †

Honorary Vice Presidents:

A. Abramson	P. Janota †	H. Morioka	M. Shirt
S. Agrawal	W. Jassem	R. Nasr	E. Stock
L. Bondarko	M. Kohno	T. Nikolayeva	M. Tatham
E. Emerit	E.-M. Krech	R. K. Potapova	F. Weingartner
G. Fant †	A. Marchal	M. Rossi	R. Weiss

Auditor:
Angelika Braun

Treasurer:
Ruth Huntley Bahr

Affiliated Members (Associations):

American Association of Phonetic Sciences	J. Hoit & W.S. Brown
Dutch Society of Phonetics	B. Schouten
International Association for Forensic Phonetics and Acoustics	A. Braun
Phonetic Society of Japan	I. Oshima & K. Maekawa
Polish Phonetics Association	G. Demenko

Affiliated Members (Institutes and Companies):

KayPENTAX, Lincoln Park, NJ, USA	J. Crump
Inst. for Advanced Study of the Communication Processes, University of Florida, USA	H. Hollien
Dept. of Phonetics, University of Trier, Germany	J.-P. Köster
Dept. of Phonetics, University of Helsinki, Finland	A. Iivonen
Dept. of Phonetics, University of Zürich, Switzerland	S. Schmid
Centre of Poetics and Phonetics, University of Geneva, Switzerland	S. Vater

International Society of Phonetic Sciences (ISPhS) Addresses

www.isphs.org

President:

Professor Dr. Ruth Huntley Bahr
President's Office:
Dept. of Communication Sciences
and Disorders
University of South Florida
4202 E. Fowler Ave., PCD 1017
Tampa, FL 33620-8200
USA
Tel.: ++1-813-974-3182
Fax: ++1-813-974-0822
e-mail: rbahr@usf.edu

Secretary General:

Prof. Dr. Mária Gósy
Secretary General's Office:
Kempelen Farkas Speech
Research Laboratory,
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
++36 (1) 321-4830 ext. 172
++36 (1) 322-9297
e-mail: gosy.maria@nytud.mta.hu

Guest Editors:

Dr. Katarzyna Klessa
Guest Editor's Office:
Department of Phonetics
Institute of Linguistics
Department of Modern Languages
and Literature
Adam Mikiewicz University in
Poznań
Sekretariat IJ UAM
Collegium Novum
al. Niepodległości 4, pok. 218 B
Poland
Tel.: ++48-829-36-63
e-mail: klessa@amu.edu.pl
personal website: katarzyna.klessa.pl

Dr. Brigitte Bigi
Guest Editor's Office:
Laboratoire Parole et Langage,
CNRS,
Aix-Marseille Université
5 avenue Pasteur,
13100 Aix-en-Provence,
France
Tel.: ++33-413 552709
e-mail: brigitte.bigi@lpl-aix.fr

Review Editor:

Prof. Dr. Judith Rosenhouse
Review Editor's Office:
Swantech Ltd.
9 Kidron St.,
Haifa 3446310
Israel
Tel.: ++972-4-8235546
Fax: ++972-4-8122461
e-mail: judith@swantech.co.il

Cover designed by Wojciech Klessa

Technical Editor:

Dr. Tekla Etelka Gráczi
Technical Editor's Office:
Kempelen Farkas Speech Research
Laboratory,
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
Tel.: ++36 (1) 321-4830 ext. 171
Fax: ++36 (1) 322-9297
e-mail: graczi.tekla.etelka@nytud.mta.hu

INTRODUCING THE GUEST EDITORS

Dr Katarzyna Klessa. Her interests focus on the analysis and development of spoken language resources, especially with application to speech prosody. In 2002, she participated in the process of creation and analysis of the *PoInt* corpus of quasi-spontaneous dialogues. In 2006 she defended her PhD on the analysis of segmental duration for the needs of Polish speech synthesis (at Adam Mickiewicz University in Poznań). In the years 2006-2010, she was involved in research-development projects aiming at creating very large text and speech corpora for automatic speech and speaker recognition for Polish. From 2011, following her interests in various kinds of speech and language resources, she has become involved in several projects devoted to the development of endangered languages archives and dissemination of knowledge (see e.g.: languagesindanger.eu). In 2012, she coordinated the design and development of *Paralingua*, a corpus for the study of linguistic and paralinguistic features in Polish, including multi-channel recordings of conversational speech, and emotion portrayals. In 2013, she has initiated the design and development of *Annotation Pro*, a freely available software tool (annotationpro.org) for annotation of linguistic and paralinguistic features in speech. *Annotation Pro* enables multilayer annotation of speech recordings using both discrete and continuous rating scales, as well as automatic annotation mining. The functionality of the programme can be flexibly extended thanks to plugin architecture. A number of plugins have so far been created and made publicly available (annotationpro.org/plugins), e.g. plugins for timing relationships analyses, such as Annotation Pro + TGA, SRMA (Speaking Rate Moving Average), and others. *Annotation Pro* is characterized by high interoperability because it offers a wide range of import/export options from/to most of the existing speech annotation tools.

Dr Brigitte Bigi. From 1997 to 2000, she worked with Professor Renato De Mori at LIA, France. She worked on statistical language modelling for automatic speech recognition and information retrieval. She has introduced a new effective model for topic identification. From 2000 to 2002, she worked with Professor Jean-Paul Haton and Pr Kamel Smaili at LORIA, Nancy, France. Her work focused on topic identification in newspaper articles and e-mails. From 2002 to 2009, she worked at LIG on statistical language modelling for automatic speech recognition and statistical machine translation. Since 2009, at LPL (Laboratoire Parole et Langage, Aix-en-Provence, France), her research has focused on corpus creation and annotation of speech recordings. The main problem she is interested in is to automatically time-

align speech data with textual data and to exploit the time-aligned results. Her research focuses on language-independent approaches to tools and systems development so that they can be used either for languages with few available data resources or for languages with unexpected amount of – unnecessary – data. She is the author and developer of SPPAS: Automatic Annotation of Speech, which includes 7 automatic annotation components (Momel and INTSINT, IPUs-segmentation, Tokenization, Phonetization, Forced-Alignement, Syllabification, and Repetitions detection), and 6 components for the analysis of annotated data.

EDITORIAL NOTE: TOOLS FOR PHONETICS

Phonetics is one of the fields of linguistics where various tools and devices have always been welcomed as useful support both for data preparation and analysis. Today, in the Internet era, phoneticians (and anyone else interested in speech science) have a great number of software tools to choose from and consequently, often need to make difficult choices. The number and variety of tools offers challenges not only to the users, but also to the software designers and developers. The challenges are related to e.g., the need for intersystem operability, re-usability, the choice of underlying methodologies, different ways of sharing the tools, as well as effective communication between software designers, developers, and the end-users. Considering the above, and following our own great interest in the field, we have decided to dedicate the present issue of The Phonetician to various aspects and perspectives of phonetics software design, development, and to describe how these software tools can be used for phonetic research.

The motivation and aims behind the paper selection for the volume were twofold. First, we wished to present a variety of freely available tools useful for the investigation of different phonetic phenomena in various languages and speech styles, and provide examples of how to use these tools. Another important issue was the need to initiate a more general discussion of the possible perspectives and future scenarios for the area of tool use and development.

The authors of the articles included in this volume contribute to this discussion by demonstrating their own tools “in action”, pointing out critical issues with these particular tools and noting more general problems such as the interface between software development and research methodology, research workflow, tool applicability and more technical questions of software accessibility and the choice of technology platforms. Dafydd Gibbon and Jue Yu discuss the methodology and implementation behind the Time Group Analyzer (TGA), a recently created on-line tool enabling a wide range of duration-based analyses. This software tool provides a broader context for the investigation of timing variability in spoken utterances. Mietta Lennes and colleagues compare pitch distributions using newly developed scripts for Praat and R for the study of intra- and inter-speaker pitch differences under various conditions. Brigitte Bigi presents SPPAS, a tool for automatic annotation and analysis of speech data, as a part of a proposed multilevel corpus creation and annotation workflow. Mark Huckvale demonstrates web audio techniques and applications, and draws attention to the recent technological change caused by the increasing prevalence of new portable platforms as opposed to the use of conventional computers for many automated tasks, including speech analysis.

We believe that by presenting this collection of articles to the readers of The Phonetician, we will achieve at least some of the assumed goals and will contribute to the discussion of the possible perspectives and scenarios in the development and use of software tools. Surely, these few papers could cover only several selected issues but they raise a number of important points. We are convinced that as a whole, the

volume provides information that can effectively support research studies and yield much food for thought.

The editors wish to thank all of the authors not only for contributing their work to this issue of The Phonetician, but also for sharing their opinions and views on this content area. We also thank the reviewers for their valuable comments, suggestions, and lively discussions.

Katarzyna Klessa and Brigitte Bigi

the Phonetician

A **Peer-Reviewed** Journal of ISPhS/**International Society of Phonetic Sciences**

ISSN 0741-6164

Number 111-112 / 2015-I-II

CONTENTS

TIME GROUP ANALYZER: METHODOLOGY AND IMPLEMENTATION

Dafydd Gibbon[1], and Jue Yu[2]
[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
Bielefeld, Germany
[2]School of Foreign Languages, Tongji University
Shanghai, China
e-mail: gibbon@uni-bielefeld.de, erinyu@126.com

Abstract

The *TGA* (*Time Group Analyser*) tool provides efficient ubiquitous web-based time-saving computational support for phoneticians without computational skills or facilities who are interested in selected linguistic phonetic aspects of speech timing. The input module extracts a specified tier (e.g. phone, syllable, foot) from a single annotation file in the common Praat TextGrid and CSV formats; user-defined settings permit selection of sub-sequences such as inter-pausal groups, and thresholds for minimum duration differences. Several types of output are provided: (1) Tabular outputs with descriptive statistics (including dispersion models like *standard deviation*, *PIM*, *PFD*, *nPVI*, *rPVI*), *linear regression*; (2) novel visual information about duration patterns, including difference *n-grams* and *Time Trees* (temporal parse trees); (3) graphs of duration relations, including Wagner Quadrant graphs. Examples of applications in phonetics are taken from published studies of varieties of Mandarin and English as a form of functional field evaluation of the tool. Other disciplines in which duration analysis has practical uses, such as forensic phonetics, clinical linguistics, dialectometry, speech genre stylometry and language acquisition, will also benefit from the efficient methodology provided by the TGA.

Keywords: online tools, speech timing, speech prosody, annotation processing, duration, time trees

1 Problems, methods, tools, solutions

1.1 Background and overview

Scientific methods are recipes for creating solutions to problems, and the tools used within these methods are the utensils which are used to implement these recipes. The tools themselves embody further methods: for example, in phonetics and speech technology, a descriptive and modelling methodology use annotations, i.e. the pairing of sections of a transcription with sections of a speech signal by means of time-stamps. The annotation procedure requires further methodological assumptions: first, on the categorial perception of speech (in creating the transcription), and second, on the physical parameters of digitized speech signals (in assigning time stamps which point to boundaries or peak points in the recorded signal).

In the history of phonetics, annotation methods have progressed from the traditional 'impressionistic' transcription of perceived sounds in an observed utterance, through recordings using various techniques and manipulation of these recordings. Before the advent of PCs such methods were common outside specialized phonetics labs, and until recently were common among phoneticians in less affluent regions. The concept of speech signal annotation arose with the speech technologies in the 1970s and software such as esps/Waves appeared in the 1980s, for the purpose of searching for, identifying and classifying portions of the speech signal in order to develop speech recognition and synthesis models. The technique of annotation was largely unknown outside of this field until free and public domain software with graphical user interfaces, such as Praat, Transcriber, Wavesurfer became available, starting in the 1990s. Newer annotation software designs with additional analysis facilities are still appearing in the interests of increased functionality and efficiency (e.g. in this volume: Annotation Pro, with facilities for perception experiments, and SPPAS, with automatic annotation based on dictionaries and statistical segment models). Annotation software supports the annotation process (1) by providing measurements and visualizations of various models of the speech signal, such as amplitude and energy envelopes, spectrum, fundamental frequency, and (2) visualizations of the mapping of arbitrarily many layers (tiers) of transcriptions and linguistic categories to segments of the speech signal. Some of these annotation tools such as Praat provide scripting languages which support the automation of particular measurement and analysis procedures. Some of the tools contain functions for exporting data and results in formats suitable for further analysis by means of other software such as spreadsheets or using modern programming languages such as Python or the statistical programming language R. An intermediate stage is represented by tools with scripting languages (e.g. the Praat scripting language) which can capture typical 'recipes' of analysis sequences, record them as scripts, and execute the scripts to analyse speech recordings automatically.

Although programming techniques are well known and widely used in specialized phonetics labs and research departments, there are still many phoneticians world-wide who use the phonetic tools such as Praat for manual numerical analysis, but lack programming skills or helpers, and are not familiar with the technique of annotation and annotation based analysis. Consequently, 'low tech' methods, for example copying on-screen values of signal parameters, such as temporal information, to spreadsheets for further calculation, are still very widespread.

The TGA online[1] 'multitool' described in this contribution is a little different, and is intended to fill a gap for the 'ordinary working phonetician' who is interested in aspects of speech timing and has no or little experience of programming. The TGA tool is is a 'multitool' in the sense that it puts together a broad set of procedures for analysing speech timing, some well known and some new, and produces a variety of

[1]Current URL: http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

analyses of timing relations in a single 'one-click' tool. An offline prototype of the TGA tool exists for handling larger amounts of data.[2]

The TGA tool itself, combining previous separate tools, was originally developed for small projects and for phonetics teaching, in a cooperation between the authors of the present contribution for the description of timing in Mandarin, in dialects of Chinese, and in Chinese English (Yu & Gibbon, 2012, Yu et al., 2014, 2015; Gibbon, 2013; Yu, 2013). The TGA tool in its current online form is designed for the analysis of timing relations in single annotation tiers from single annotation files. Timing analyses across more than one tier are not incorporated in the present version; if such analyses are required the separate results must be exported, combined and further analysed with a spreadsheet or other software application.

The online TGA user interface design is kept very simple: an annotation file is opened in a text editor, copied and pasted into an HTML form on a web page. Parameter settings permit the selection of the relevant tier name and values of parameters for the analysis, and a 'one-click' timing analysis takes place, using a range of analysis procedures and based on the time-stamps in the data, and producing a wide variety of outputs (see Section 3). In addition to common measures such as speech rate and variability, similarity or dispersion of duration values (e.g. *standard deviation*, *nPVI*, described in Section 2), novel measures and displays of acceleration, visualization of regularities by bar charts, time function plots and scatter plots are included, as well as chracter separated value (CSV) outputs for further analysis with statistical tools. A preliminary version of the TGA has been previously described (Gibbon, 2013). Components of the TGA tool have been incorporated in software by other developers (AnnotationPro and SPPAS, this volume). Typically, TGA applications have been applied to the syllable tier, but the duration of intervals on any tier in an annotation can be selected and analysed.

The objective of this contribution is to provide an account of TGA tool development strategy from problem domain through specifications to implementation. It is not primarily a manual for how to use the tool for a specific phonetic analysis purpose, though application examples are given in Section 3.4.

The organization of the contribution follows a general scheme covering problems, methods, tools, solutions, roughly according to a traditional software development procedure of requirements specification, design, implementation and evaluation. The following subsection 1.2 delimits and characterizes a selection of problems in syllable duration analysis. Section 2 deals with a set of linguistic phonetic methods which have been proposed for solving the problems, and with new methods for new aspects of the domain. In Section 3 specification, design, implementation and phonetic applications of the TGA tool are described, as well as its application in selected publications on

[2]There is an offline development prototype capable of handling larger amounts of annotation data and with additional functions which are not available online owing to server limitations. The offline prototype is not yet available for general distribution.

timing problems. Section 4 concludes the description by outlining areas which have been addressed with the TGA, and by addressing planned extensions and noting practical application potential in neighbouring disciplines.

1.2 Aspects of speech timing: delimiting the TGA domain

The domain of issues handled by the TGA tool is characterized in Section 2 The aim of this subsection is simply to delimit this domain very briefly in the context of a broader range of issues in subsegmental, segmental and suprasegmental or prosodic speech timing, ranging from voice onset time and stop closure-opening time through vowel, consonant and syllable reduction, speech rate and rhythm through pause patterning to timing in discourse. Figure 1 compactly summarizes the rank-interpretation hierarchy of the language structures, functions and phonetic correlates involved. The TGA tool focuses on sequential and hierarchical relations within sequences of units such as phones, syllables, words (depending on the annotation tier selected). The TGA tool is in principle suited to analysis of units at any level of the rank-interpretation hierarchy shown in Figure 1, but has so far been mainly restricted to analyzing temporal relations between syllables in Time Groups of two kinds: (1) interpausal time groups, and (2) time groups based on acceleration and deceleration of speech rate (e.g. syllable rate).

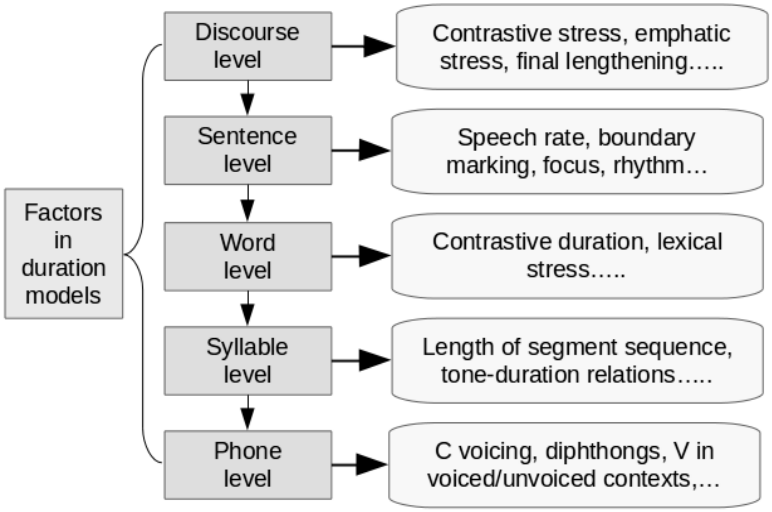


Figure 1: Domains of speech timing patterns

One of the areas of deployment of the TGA tool is in the study of aspects of speech rhythm, an area which has been conspicuous in the phonetic literature since the study of Pike (1945) on the intonation of American English. One of the questions involved has been whether and how the perception of rhythm in different languages tends towards two poles of *syllable timing* on the one hand, and *foot timing* (with related concepts such as *stress timing*, *interstress timing*) on the other. Searches for correlates of rhythm in the speech signal have been somewhat inconclusive (Arvaniti, 2009), motivating a view

that rhythm is an epiphenomenon which cannot be simply induced from the temporal patterning of physical speech signals and which results from the interplay of many factors, including those outlined in Figure 1: discourse and grammatical structure, word familiarity and frequency, morphological structure and phonotactic patterns (Gibbon, 2006). The physical correlates in turn involve several parameters: the timing of units of speech, as well as pitch and intensity patterns. Nevertheless, the search is not over, and the function of the TGA tool is to support research specifically in relation to speech timing in matters including but not limited to rhythm.

There is currently no comprehensive theory of speech rhythm production and perception, and no model of rhythm patterns. An earlier pretheoretical clarification of the term 'rhythm' was summarized by Gibbon et al. (2001) as an iteration of alternations of strong and weak values of some parameter or parameter set, whose alternations which have a tendency to isochrony. The model may be termed a 'Three Constraint Model' of rhythm:

> Rhythm is the recurrence of a perceivable temporal patterning of strongly marked (focal) values and weakly marked (non-focal) values of some parameter as constituents of a tendentially constant temporal domain (environment).

This Three Constraint Model has turned out to be inadequate in a number of respects as it is missing the similarity and hierarchy properties of speech rhythm (Gibbon, 2003), and of rhythm in music and other domains. A 'Five Constraint Model' is more adequate, requiring fulfilment of the following criteria, which will figure in the description of the TGA tool:

1. a dynamic *Alternation Constraint* on patterns of stronger and weaker elements of some parameter or parameter set;
2. an oscillatory *Iteration Constraint* on repetition of adjacent patterns;
3. a qualitative *Similarity Constraint* on elements of the iterated adjacent patterns;
4. a quantitative *Isochrony Constraint* on the iterated adjacent patterns;
5. a structural *Hierarchy Constraint* on rhythm, which specifies temporal domains in a relation of temporal inclusion, to each of which the previous constraints apply (the temporally shortest alternation being the lowest and sometimes the only level in the hierarchy).

The basic strong-weak Alternation Constraint applies at different structural levels in different languages. Typical of tendencies to the 'ideal type' of syllable timing is the alternation consonants and vowels (CV, CVC patterns), and in the 'ideal-type' of stress timing is alternations between strong syllables and one or more short syllables. The 'ideal-types' are in practice only approximative tendencies; so-called stress-timed languages may also have fortuitous syllable timing: *Jim swam fast past Jane's boat*, and vice versa.

Recent approaches (Cummins, 2009; Inden et al., 2013; Włodarczak, 2012) have addressed more complex issues of modelling rhythm by means of oscillators and of

the mutual adaptation or entrainment of rhythms by interlocutors in discourse; this domain is outside the immediate scope of the present study, and inter-tier duration relations are currently not included in the specification of the TGA tool.

2 TGA prerequisites: approaches to prosodic timing

2.1 Phonological and linguistic phonetic approaches

The present section concentrates on the aspects of speech timing analysis for which support by the TGA tool is designed. Overviews of relevant methods of timing analysis at the level of syllable patterning are given by Gibbon (2006) and the contributions to Gibbon et al. (2012). These methods presuppose some prior identification of linguistic and phonetic categories in the form of segmentations and labellings of speech recordings, whether by annotation or direct measurement of signal visualizations. Many analysis methods have been applied to the problem of examining duration relations between consonantal and vocalic syllable constituents, or between syllables, or between stress-based feet. However, most have concentrated implicitly or explicitly solely on the *Iteration Constraint* and the *Isochrony Constraint* outlined in the Section 1.2, to the exclusion of the *Alternation Constraint*, the *Similarity Constraint* and the *Hierarchy Constraint*.

Timing hierarchies have been discussed in several different theoretical and methodological contexts: in post-generative phonologies such as Metrical Phonology (Goldsmith, 1990); as prosodic structure (Jassem, 1952; Abercrombie, 1967); as oscillation (Barbosa, 2002; Inden et al., 2012). In the present contribution, novel methods for modelling the *Alternation Constraint* and the *Iteration Constraint* and the *Hierarchy Constraint* as *Duration Difference Token* (*DDT*) sequences is presented in the present contribution, and the *Time Tree* (*TT*) method of timing hierarchy induction (Gibbon, 2003, 2006) is also discussed.

The comprehensive structural rhythm model which has been most extensively investigated phonetically is that of Jassem (1952) and Jassem et al. (1984), which invokes *alternation* (of stressed syllable and sequences of unstressed syllables), *iteration* (of stressed-unstressed alternations), *similarity* and *near-isochrony* (of stressed-unstressed sequences) and *hierarchy* (of broad and narrow rhythm units). The Abercrombie model addresses the same constraints but with a simpler structure and without the hierarchy constraint. The Jassem model and to some extent the Abercrombie model (1964:219) also take morphological structure (word boundaries) into account.

In Jassem's model, the Broad Rhythm Unit (BRU) has two constituents: an optional Anacrusis (ANA), consisting of unstressed syllables from a grammatical boundary (e.g. utterance, phrase, word boundary) up to but not including the next stressed syllable, and an obligatory Narrow Rhythm Unit (NRU), consisting of a stressed syllable followed optionally by a sequence of unstressed syllables, extending to the next relevant grammatical boundary. Thus, a neutral pronunciation of the sequence *it's stressful today* may yield the following parse:

(BRU: (ANA: it's) (NRU: stress ful)) (BRU: (ANA: to) (NRU: day))

However, the better known model is the simpler and flatter model of Abercrombie (1967), who analyses sequences feet, each consisting of an 'ictus' (a phonetically stressed syllable, which may be phonemically long, medium or short) and a 'remiss' (an optional sequence of unstressed syllables). Initial sequences of unstressed syllables are treated as having an empty ictus or null beat:

|| - it's | stress ful to | day ||

Jassem et al. (1984) have shown that the more complex Jassem model fits the English facts better than the simpler Abercrombie model: no empty beat is needed, and they showed experimentally that unstressed syllables in the Anacrusis have different timing properties from those in the Narrow Rhythm Unit (cf. also contributions to Gibbon et al., 2012 for extensive discussion).

The Jassem and Abercrombie models are both very close to the present Five Constraint model of rhythm in that they incorporate the Alternation, Iteration, Similarity, and Isochrony Constraints and (in the case of the Jassem model) also the Hierarchy Constraint. These Jassem and Abercrombie models and the Five Constraint model are not explicitly included in the domain of the TGA, but need to be borne in mind when using the TGA tool for analysing the relation between phonological and phonetic determinants of speech timing, particularly rhythm.

2.2 Linear quantitative models of duration dispersion

The inclusion of a selection of linear quantitative models of duration in the domain of the TGA tool requires explicit justification. Several studies of speech timing have concentrated on subsyllabic or syllabic properties, looking at the dispersion and percentatages of consonantal and vocalic stretches of the speech signal, for example variance or standard deviation of the durations of consonantal intervals (ΔC) and percentage of vowel durations (%V). Measurements based on the ΔC–%V model introduced by Ramus et al. (1999) yielded interesting results about the differentiation of different languages by means of the relation between these parameters and between these parameters and other dispersion measures such as vocalic normalized pairwise variation (*nPVI*) and consonantal raw pairwise variation (*rPVI*); cf. Low et al. (2000) and very many studies using these two measures. A selection of approaches of this type is shown in Table 1, including two already mentioned.

The top two models in Table 1 are the *Pairwise Irregularity Model* (*PIM*) of Scott et al. (1986) which sums all pairwise log ratios of each interval duration in the whole utterance, and the *Pairwise Foot Deviation* (*PIM*) model of Roach (1982), which takes adjacent pairwise differences rather than all pairwise differences, and is rather like standard deviation, except that the absolute magnitude of differences is taken, rather than the square and the square root. Although the Roach model refers to the foot as a unit, formally speaking the models are agnostic in regard to the units to which they apply.

The bottom two models, which have already been referred to, are variants of an *Average Magnitude Difference Function* (*AMDF*), in which differences in a moving window over pairs of adjacent intervals are averaged. This results in factoring out variations in speech rate, a useful innovation. In the context of speech timing analysis,

the binary window AMDF is known as the *raw Pairwise Variability Index* (*rPVI*). The *rPVI* takes iteration and isochrony into account, but not alternation and hierarchy. The *rPVI* is normally applied to consonantal intervals in a speech recording, and is distinguished from the *normalized PVI*, *nPVI*, normally applied to vocalic intervals. The duration differences in the *nPVI* are normalized by dividing each difference by the average of the durations. The overall average is multiplied by 100 (as in the *PFD*), resulting in a scale for the *nPVI* from 0 (complete isochrony) to an asymptote of 200 (completely random).

Table 1: Four dispersion models of speech segment patterning.

PIM($I_{1,...n}$)	=	$\sum_{i \neq j} \lvert \log \frac{I_i}{I_j} \rvert$
PFD($foot_{1...n}$)	=	$\frac{100 \times \sum \mid MFL - len(foot_i) \mid}{len(foot_{1...n})}$ where MFL $= \frac{\sum_{i=1}^{n} len(foot_i)}{n}$
rPVI($d_{1...m}$)	=	$\sum_{k=1}^{m-1} \mid d_k - d_{k+1} \mid /(m-1)$
nPVI($d_{1...m}$)	=	$100 \times \sum_{k=1}^{m-1} \mid \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \mid /(m-1)$

The PVI variants have become very popular since their introduction by Low et al. (2000), have been used in very many studies and have yielded very interesting results about the dispersion of duration relations between different languages. In the literature there has been plenty of folklore and various simple misunderstandings about the nPVI formula[3]: (1) the component 'n-1' has been said to mean that the last syllable is not considered in order to factor out final lengthening, but this is false since the formula is about differences between adjacent items in a sequence, and there is always one difference less than the total number of items, and final lengthening is not factored out; (2) the factor '100' has been said to convert the result to a percentage, but this is false since the nPVI scale is 0...200, because for normalization each duration difference is divided by the average duration of the pair (sum/2) and not the by sum, which would indeed have yielded 0…100.

Critics have also pointed out that (1) essentially the same results may be obtained from phonotactic patterns without phonetic measurements (Hirst 2009); (2) similar dispersions may occur between stylistic and dialectal varieties of the same language (Gut, 2012; Arvaniti, 2009); (3) in the PVI and PFD models the pairwise differences between adjacent syllables imply that rhythms are purely binary, for example with alternations of long and short syllables. This is not necessarily the case in stress-accent

[3]These will not be cited here in order to avoid embarrassment.

timed languages, however, where several unstressed syllables may intervene between stresses.

These measures have been (and often still are) called ‘rhythm metrics’, but this is a misnomer since, like plain standard deviation, none of these four measures fulfils either the Alternation, or the Iteration, or the Hierarchy Constraint; each concentrates only on a dispersion measure for relative isochrony. This is a fundamental formal criticism. With the first two models, ordering the values in any order, whether by in order of occurrence, or in increasing or decreasing or random order, yields the same dispersion values. With the PVI models it is also possible for different patterns to yield the same value, e.g. an alternating pattern like 2 4 2 4 yields the same value as 2 4 8 16, or 2 4 8 4, namely 66.6'. The reason for this oddity is the use of absolute magnitudes (the ‘|...|’ notation) with the result is that the direction of differences or ratios becomes irrelevant and therefore the Alternation Constraint is factored out.

Another fundamental criticism which applies to all four models, is that that the nearer the index is to zero, the more similar the timing pattern is to syllable (or foot, etc., depending on the unit being measured) timing. The further away from zero the index is, the less is known about what units are actually being measured, and the less one can be certain about whether it is a rhythm which is being measured (Gibbon, 2003). It is thus impossible to know what these results actually mean without combining them with further studies of units of different size, and taking the alternation, iteration and hierarchy constraints into accounts. The models account for a subset of the necessary conditions for rhythm, but do not provide a sufficient condition.

However, as measures of smoothness, regularity or relative isochrony relative to a unit such as a consonantal, vocalic, syllabic or foot interval the measures yield consistently useful results in demonstrating differences between languages. Examples of such analyses obtained with the TGA tool will be given in the case studies of applications in Section 3.4.

2.3 Dynamic timing factors: speed and acceleration

Values such as the minimum and maximum values of interval durations are subject to large fluctuations determined by the wide range of determining factors shown in Figure 1. However, useful further notions are connected with the speed of speech, usually measured in terms of the rate of phonemes, syllables, feet, stresses, phrases, etc. per second. The rate is the inverses of the mean duration; this, if the mean syllable length is 125 msec, the syllable rate is 8 syll/sec.

Another interesting parameter is the rate of change of speed, i.e. the overall acceleration or deceleration of a sequence of units such as the syllable, whether very locally with long-short syllable pairs or over an entire utterance. If measured over a long sequence, a useful measure is provided by linear regression models: the resulting slope indicates acceleration (if negative, i.e. with decreasing interval durations) and deceleration (if positive, i.e. with increasing interval durations).

These measures of speed and acceleration-deceleration are included in the TGA tool, and examples of the use of these measures are discussed in the application case studies in Section 3..4.

2.4 Patterns and relations: data visualization

An important part of scientific methodology, both with experimentation on small data sets and with inductive analytics applied to 'big data' is the visualization of data structures and distributions as a source of insights for explanations. Particularly useful visualizations of speech timing data have been forthcoming from use of the ΔC–%V model and the *PVI* models, sometimes in combination, in illustrating similarity clusterings and differences among languages, as previously discussed.

There are other forms of visualization which can be very helpful. Even a straightforward plot of durations as a function of time enables an instant intuitive assessment of temporal evenness or variability (see the *Implementation* section of this contribution). Even more useful is the Wagner Quadrant visualization method (Wagner, 2007) for showing the relations between adjacent interval durations without using the absolute magnitude, a method which was developed as part of a criticism of the methods shown in Table 1 and discussed above, and which, unlike those measures, does not factor out the directionality of differences.

Sections 3.3.2 and 3.4 provide examples of the different kinds of visualizations which are provided in the TGA output.

3 The Time Group Analyzer (TGA) tool

3.1 TGA Requirements specification

As noted previously, methods are recipes for creating solutions to problems, tools are the utensils which are used to implement these recipes, and each utensil is itself based on other theories and models. Practical recipes for the analysis of speech sounds have been around for a long time, and software timing analysis tools may be seen as the utensils for these recipes. There are many other kinds of tools. For example, teachers of English as a foreign language know about 'gesture tools' such as the dodge of isochronous tapping on the table and clapping or drumming rhythmically in time with stress beats (though these rhythms may be far from the properties of natural live English speech). A variant of the same isochronous tapping has been the use of a metronome tool in experimental work on timing entrainment (Cummins, 2009).

The TGA tool exploits each of the four main steps involved in creating the input annotations:

1. Extraction of the relevant annotation tier, representing an attribute (i.e. feature type)
2. Extraction of the text of the tier, i.e. the values of the attribute represented in the tier (e.g. phonemes, syllables, feet, phrases, tones stresses, boundaries); currently a subset of UTF-8 encoding is handled, but X-SAMPA encodings, rather than IPA glyph codes are preferred.
3. Extraction of the time-stamps representing association of the sequence values represented in the tier with segments of the signal.
4. Analysis and visualisation of information derived from the time-stamps.

Thus the annotation process essentially follows the segmentation and classification procedures of structuralist phonetics and phonology, and the TGA tool picks up the

thread at this point by analyzing temporal relations between time-stamped segments in the selected tier. Input formats for annotations are Praat TextGrids in long or short format, or Character Separated Value (CSV) tables. Other annotation tools than Praat such as Elan and Annotation Pro or SPPAS have import and export functions for these formats as well as their own formats. None of these formats is particularly complex and it is fairly simple to convert one into another. TGA analyses identify the speech rate of the segments on the selected tier such as phones or syllables, duration dispersion by standard deviation and previously mentioned similar functions which yield measures of relative, 'sloppy' or 'fuzzy' near-isochrony, either relative to adjacent units (e.g. *rPVI*, *nPVI*), or relative to the whole sequence, as with standard deviation, the *PIM*, and the *PFD*.

3.2 TGA design

The literature reveals several common methods for processing time-stamped data, in order of increasing sophistication:

1. copying into spreadsheets, sometimes using templates available on the internet for semi-manual processing: a traditional procedure, still common outside well-equipped labs and phonetics departments;
2. use of online tools for specific purposes, such as *nPVI* or speech rate calculation, and further processing with spreadsheets or specialised statistics software;
3. use of prefabricated or *ad hoc* Praat scripts to create numerical output for further processing;
4. implementation of applications in appropriate scripting languages such as *Perl*, *Tcl*, *Ruby*, *R* or *Python*;
5. implementation in languages such as *C*, *C++*, mainly in specialised speech technology applications), independently of time-stamping visualization software.

The TGA online tool falls into the second of these classes, thus filling a gap between non-programming and programming approaches, within a circumscribed functionality for duration analysis, and side-stepping the need for the 'ordinary working phonetician' to use programming techniques. For those with programming abilities, libraries of analysis tools are available, e.g. those in *Perl* in the Aix-MARSEC repository (Auran et al., 2004), or parsing functions programmed in *Python*, such as the *Natural Language Took Kit*, *NLTK* (Bird et al., 2009), or the *TextGrid tools* (Buschmeier et al., 2013).

The architecture of the TGA online tool is shown in Figure 2. Input from an HTML form is passed to a server on the internet (or a localhost server on a standalone machine) and processed by a number of TGA modules, with a variety of output types. The basic design is heavily dependent on the theoretical assumptions outlined in Section 2.

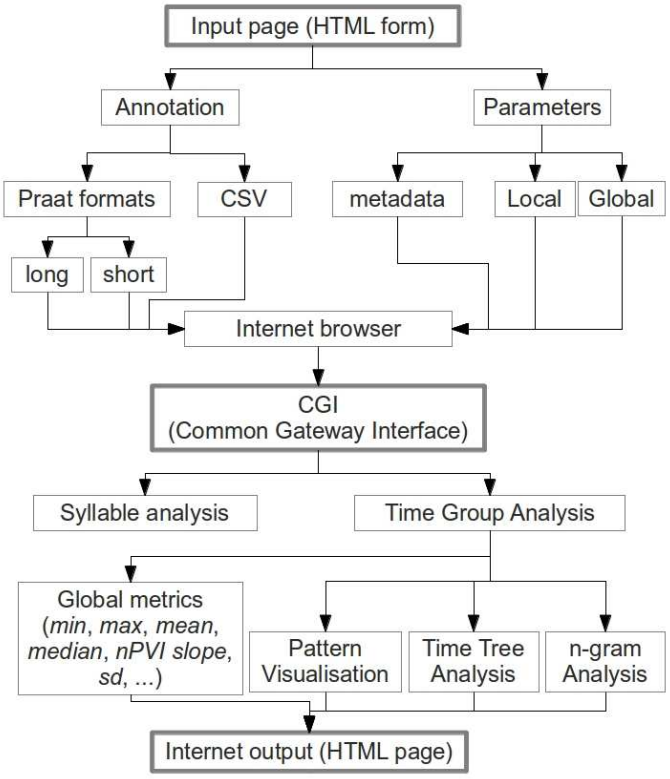


Figure 2: Online TGA architecture.

3.3 TGA Implementation

3.3.1 Input format and parameter setting. The TGA tool is currently implemented in Python 2.7 as a server-side application in the CGI internet environment. The choice of an online environment has many advantages: operation in a standard browser; consistent (because identical) environment at any given time. A disadvantage which is sometimes mentioned is that data input into online tools may be collected on the server by the tool provider. This does not happen with the TGA; user data are neither inspected nor ollected, and user anonymity is preserved.

Input identification and parameter setting in the TGA tool are shown in Figure 3. The parameters are organized into three functionally related groups: input identification, processing parameters, and output selections.

The TGA input module extracts a specified tier (e.g. phone, syllable, foot) from inputs in long or short TextGrid format, or as character separated value (CSV) tables with any common separator. The example specifies a tier ‘Syllables’ and a set of pause symbols which may be used. The pause symbols may be freely selected as long as they do not clash with names of other text labels. The underscore ‘_’ shown in the figure is a very common pause symbol.

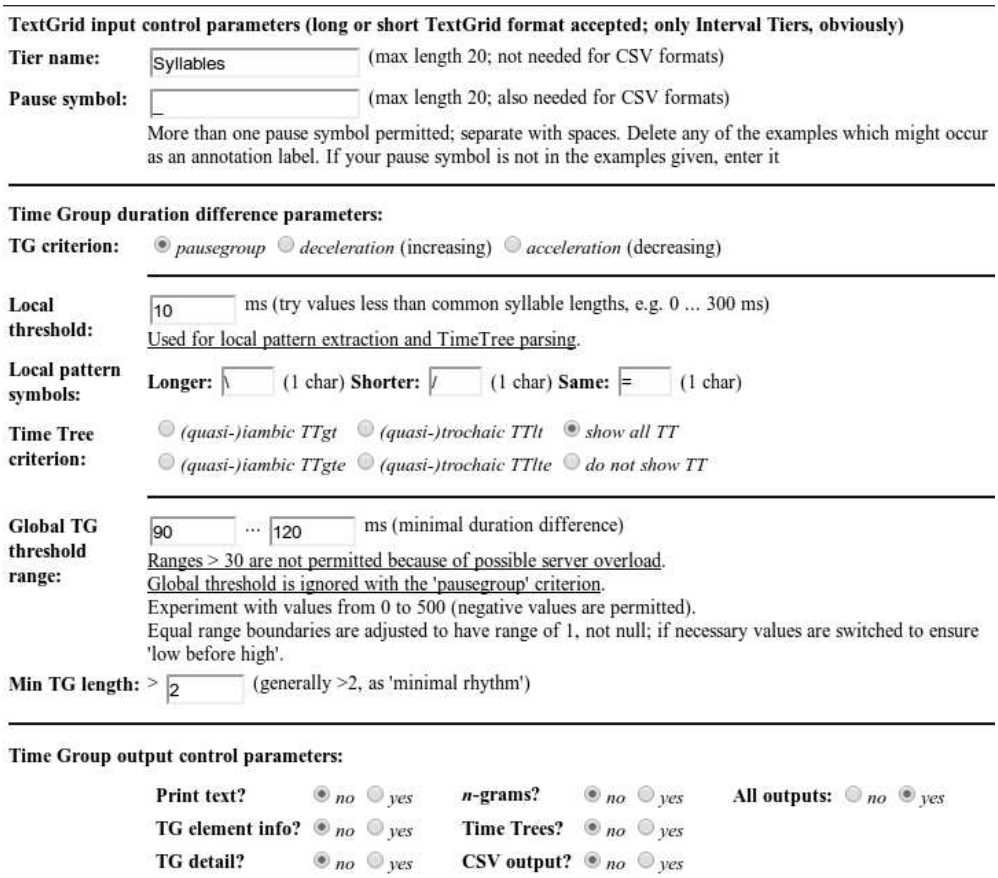


Figure 3: Screenshot of parameter input options.

In the processing parameter section, the local threshold permits specification of the minimal difference in milliseconds between durations which determines which durations count as different and which count as equal. The local threshold is relevant for constructing the Duration Difference Tokens (DDTs) described in Section 3.3.2.3 and the Time Trees (TTs) described in Section 3.3.2.4: the larger the threshold, the more duration pairs count as equal, removing random 'duration difference noise'. The DDT symbols can be freely defined. Four TT types are defined, two based on short-long pairs (quasi-iambic, pairwise deceleration), two based on strong-weak pairs (quasi-trochaic, pairwise acceleration).

The global threshold range is a tentative experimental feature for identifying Time Groups by means of accelerating or decelerating sequences within the specified range.

The minimum Time Group length permits restriction of analysis to Time Groups with a length which promises useful numerical results.

Finally, the output parameter section specifies output of selected results from the modules (see Section 3.3.2) or of all possible outputs.

3.3.2 TGA solutions: the main modules. Currently there are three main TGA modules besides I/O and format conversion: (1) text extraction; (2) global basic descriptive statistics for all elements of the specified tier; (3) segmentation of the tier into *Time Groups* with statistics for individual *Time Groups*, and (4) three new visualization techniques for Δ*dur* duration patterns: duration difference tokens, duration column charts, and *Time Trees*.

3.3.2.1 Text extraction. When the annotation has been made directly with annotation software, without prior transcription, there may be a need for transcription text extraction, as documented by a number of web pages providing this functionality, for various purposes such as discourse analysis, natural language processing, archive search, re-use as prompts in new recordings. This facility is provided by extracting labels from annotation elements as running text, separated into sequences by the boundary criteria, e.g. pause, specified in the input. The following example of interpausal groups is extracted from an annotated recording in the CASS corpus of Mandarin (Li et al., 2000):

bei3 feng1 gen1 tai4 yang2 p
you3 yi4 hui2 p
bei3 feng1 gen1 tai4 yang2 zai4 nar4 zheng1 lun4 shui2 de5 ben3 shi5 da4 p
zheng1 lai2 zheng1 qu4 jiu4 shi4 fen1 bu4 chu1 gao1 di1 lai2 p
zhe4 shi2 hou5 lu4 shang5 lai2 le5 ge4 zou3 daor4 de5 p
ta1 shen1 shang5 chuan1 zhe5 jian4 hou4 da4 yi1 p
ta1 men5 lia3 jiu4 shuo1 hao3 le5 p
shui2 neng2 xian1 jiao4 zhe4 ge5 zou3 daor4 de5 tuo1 xia4 ta1 de5 hou4
da4 yi1 p
jiu4 suan4 shui2 de5 ben3 shi5 da4 p
bei3 feng1 jiu4 shi3 jinr4 de5 gua1 qi3 lai2 le5 p
bu2 guo4 p
ta1 yue4 shi4 gua1 de5 li4 hai5 p
na4 ge5 zou3 daor4 de5 p
ba3 da4 yi1 guo3 de5 yue4 jin3 p
hou4 lai2 bei3 feng1 mei2 far3 le5 p
zhi3 hao3 jiu4 suan4 le5 p
guo4 le5 yi2 huir4 p
tai4 yang2 chu1 lai5 le5 p
ta1 huo3 la4 la4 de5 yi2 shai4 p
na4 ge5 zou3 daor4 de5 ma3 shang4 jiu4 ba3 na4 jian4 hou4 da4 yi1 tuo1
xia4 lai2 le5 p
zhe4 xiar4 bei3 feng1 zhi3 hao3 cheng2 ren4 p
ta1 men5 lia3 dang1 zhong1 hai2 shi5 tai4 yang2 de5 ben3 shi5 da4 p

Further analysis of the text output (frequency lists of items, concordance) is planned in future versions of the TGA tool.

3.3.2.2 Global and local descriptive statistics for all Time Groups in the annotation. For calculating global descriptive statistics, three versions of the data are prepared: (1) with all annotation elements on the tier, including boundary elements (e.g. pauses); (2) with only non-boundary elements; (3) with only boundary elements; cf. Figure 4.

Duration props (syllables)

Attributes	Values	Attributes	Values
n:	275	intercept:	156.122
min:	20	slope:	0.148
max:	990	std:	113.584
mean:	176.38	nPVI:	62
total:	48504	rPVI:	10920
range:	970	-	

Figure 4: Screenshot of summary of collated *Time Group* properties and correlations.

Basic statistics, and additionally linear regression (slope and intercept) to show acceleration/deceleration, are also tabulated for each Time Group separately (cf. Table 2, with a selection). The full table output contains not only descriptive statistics for each Time Group row, as shown in Table 2, but also additional information on each row (for this cf. Figure 5, Figure 6). Some of this additional information is dependent on the setting of the minimal difference threshold parameter, which defines degrees of approximate (i.e. 'fuzzy' isochrony), rather than strict time-stamp differences. In addition to the numerical output, three novel structural Δ*dur* pattern visualizations are defined (cf. also Figure 5):

1. tokenization of duration differences Δ*dur* into 'longer', 'shorter' and 'equal' duration difference tokens, represented by character symbols (cf. Figure 5), to support prediction of whether specific properties such as rhythmic alternation are likely to make sense (threshold dependent);
2. top-suspended bar chart illustrating the duration Δ*t* of elements in the Time Group (Figure 5), the Duration Bar Sequence (DBS);
3. duration parse tree (*Time Tree*, *TT*) for each Time Group (Figure 6), based on signed duration differences Δdur^{+} and Δdur^{-} (Gibbon, 2003, 2006) to facilitate study of correspondences between duration hierarchies and grammatical hierarchies (threshold dependent).

3.3.2.3 Duration Bar Sequences (DBS) and Duration Difference Tokens (DDT). In Figure 5 two of the novel visualizations are displayed. The hanging Duration Bar Sequence (DBS) provides an iconic representation of syllable (or other selected unit) durations both in width and in height. The row of slashes above the DBS shows the directionality – i.e. alternation – of syllable duration differences as Duration Difference

Tokens (DDT). Comparison with the DBS shows that '\' represents a short-long relation, or deceleration (rallentando, iambic), and '/' represents a long-short relation, i.e. acceleration (accelerando, trochaic), while '=' represents equality of duration (depending on the currently defined local duration difference threshold). In the Mandarin example (top) the DBS shows no obvious alternation of syllables into larger structures such as feet, while the English example (bottom) shows a conspicuous tendency to alternation between long and short syllables. The DDTs show an effect of the local difference threshold: differences <= 10 ms are shown as equal. A selective distributional analysis of bigram DDT sequences is shown in Table 3, providing an indication of the degree of (binary) alternations vs. non-alternations.

Table 2: Selection of output table of local measures for each interpausal Time Group. The full table contains additional columns on the right with the transcription of the TG and visualizations on each row (cf. Figure 5). Number of Time Groups: 23 ; Total duration (without pauses): 31771 ms.

#	n	dur(ms)	rate	mean	median	stdev	nPVI	mednPVI	PIM	PFD	intercept	slope
01	00	0000	0.00	000.00	000.00	00.00	00	00	000	00	000.00	-000.00
02	05	1199	4.17	239.80	250.00	42.29	33	36	005	15	245.60	00-2.89
03	03	0531	5.65	177.00	110.00	94.75	48	48	004	50	076.50	-100.50
04	14	2516	5.56	179.71	186.00	50.48	42	39	070	22	196.11	00-2.51
05	12	1991	6.03	165.92	163.00	58.28	50	46	063	28	166.63	00-0.12
06	11	1834	6.00	166.73	161.00	54.55	34	27	049	27	154.95	-002.35
07	09	1572	5.73	174.67	173.00	52.75	26	22	026	20	135.93	-009.68
08	07	1185	5.91	169.29	181.00	50.69	55	55	018	25	143.46	-008.61
09	16	2470	6.48	154.38	153.00	53.59	40	34	108	27	138.49	-002.12
10	07	1143	6.12	163.29	181.00	50.41	54	55	019	26	167.14	00-1.28
11	10	1752	5.71	175.20	172.50	55.19	39	32	037	24	227.40	0-11.59
12	02	0371	5.39	185.50	185.50	60.50	65	65	001	33	125.00	-121.00
13	07	1149	6.09	164.14	182.00	70.25	58	56	024	36	112.50	-017.21
14	05	0876	5.71	175.20	168.00	55.76	49	52	009	24	130.00	-022.60
15	07	1218	5.75	174.00	162.00	48.33	38	38	014	22	146.89	-009.04
16	07	1332	5.26	190.29	213.00	43.60	27	32	013	21	149.57	-013.57
17	05	0935	5.35	187.00	186.00	65.08	53	54	010	28	207.40	0-10.19
18	04	0641	6.24	160.25	127.00	85.34	56	55	008	44	099.20	-040.70
19	05	0872	5.73	174.40	166.00	16.18	14	16	002	09	185.20	00-5.39
20	07	1344	5.21	192.00	169.00	81.79	42	34	022	36	191.14	00-0.29
21	18	3051	5.90	169.50	167.50	47.11	25	17	109	22	176.53	00-0.82
22	08	1557	5.14	194.63	173.50	41.86	24	19	014	19	167.92	00-7.63
23	13	2232	5.82	171.69	171.00	76.06	63	68	094	35	179.80	00-1.34

In this instance of 'educated Southern British' pronunciation, i.e. slightly modified Received Pronunciation (RP), alternations figure at the top two ranks, totalling 42% of the digrams, and therefore have potential for identification as satisfying the rhythmic Alternation Constraint; deceleration patterns (short-long relations) occupy rank 3. Analyses with thresholds higher than 10ms are necessary for more information about the Alternation Constraint (see Section 3.4.2).

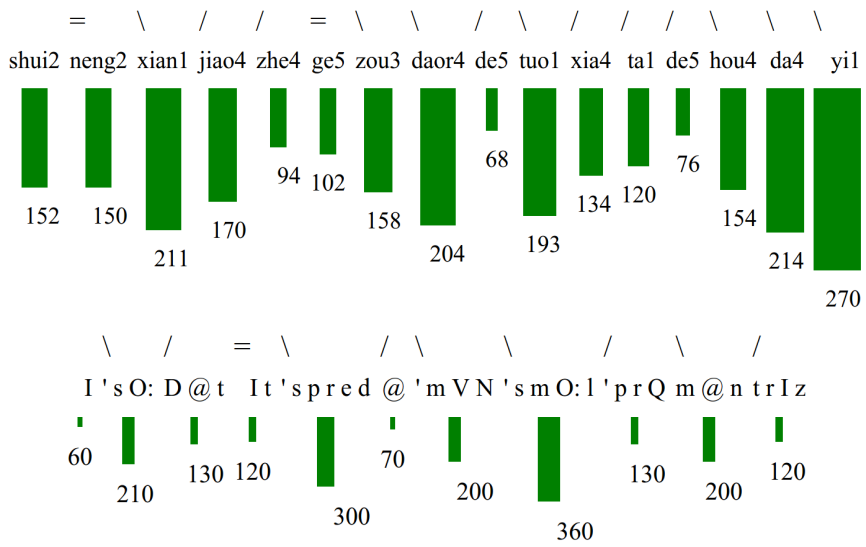


Figure 5: Top: Mandarin. Bottom: English. Duration Difference Token sequence (above) and top-suspended Duration Bars (below); duration is represented by both width and length; scaling is dependent on length of syllables in the transcription.

*Table 3: Δ*dur *t*oken rank and frequency analysis.

Rank	Percent	Count	Token digram
1	22%	60	/ \
2	20%	55	\ /
3	11%	31	\ \

3.3.2.4 Time Trees. A further non-traditional visualisation is the *Time Tree* (Gibbon 2003), which groups items in *Time Groups* into binary trees based on the alternation properties of syllables. The *Time Tree* induction algorithm follows a deterministic context-free bottom-up left-right shift-reduce parser schedule. The grammars use Δdur^{+} and Δdur^{-} tests on annotation events in order to induce two types of *Time Tree*, with 'quasi-iambic' (decelerating, rallentando) constituents, and 'quasi-trochaic' (accelerating, accelerando) constituents, whereby larger constituents inherit the longest duration of their smaller constituents. In Figure 6, a Time Tree constructed over the inter-pausal group 'about Anglican ambivalence to the British Council of Churches' is shown in nested parenthesis notation. The example is taken from the Aix-MarSec English corpus (Auran et al., 2004).

The purpose of generating Time Tree output is to support study of the relation between temporal hierarchical structures and grammatical constituents in a systematic *a posteriori* manner, rather than simply looking for timing correlates of higher level units such as feet or other event types in an *a priori* prosodic hierarchy framework. The example in Figure 6 shows a number of correspondences with grammatical units at different depths of embedding, e.g. 'about', 'British', 'Anglican ambivalence', 'about

Anglican ambivalence', 'Council of Churches', 'to the British Council of Churches', including foot sequences of Jassem's 'Anacrusis + Narrow Rhythm Unit' type.

```
( ( (@ baUt)
    ( ( ({N glI)
        (kn {m))
      (bI vl@ns)))
  ( ( ( (t@ D@)
        (brI tIS))
      ( ( kaUn
          ( sl
            (@v tS3:)))
        tSIz))
    PAUSE))
```

Figure 6: Automatic prettyprint of a quasi-iambic *Time Tree* in nested parenthesis notation.

Crucially, Δ*dur* token patterns and *Time Trees*, (unlike *standard deviation, PIM, PFD, rPVI, nPVI*) use signed, not unsigned duration differences, and may therefore claim to represent true rhythm properties. In each case, the minimal local difference threshold setting applies, determining the degree of 'fuzziness' in the distance measurement used in representing duration relations.

A detailed summary chart of the overall statistics is given in Figure 7. The numerical informtion in the chart contains averages over the individual Time Groups, and also provides correlations between the different measures.

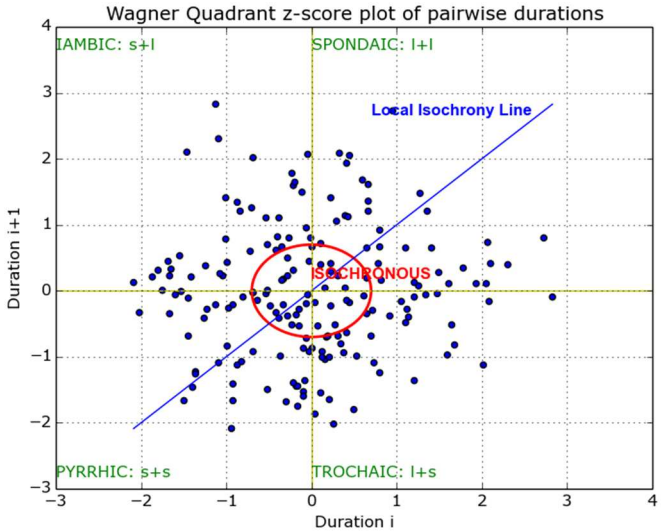
3.3.2.5 Wagner Quadrant Graphs. The main further visualization provided by the TGA is the Wagner Quadrant Graph (Wagner, 2007), a scatter plot which reflects the signed z-scores of duration differences rather than the absolute magnitude of differences. The signed differences and z-scores, i.e. (*meanduration – duration*) / *standard deviation*, were used in order to preserve comparability of data, in the context of a critique of the PVI model, which uses absolute magnitudes and raw data. The scatter plots show the duration z-scores of adjacent syllables on the X and Y axes (cf. Figure 8).

The differences between Mandarin and English syllable duration dispersions are shown very clearly. Mandarin syllable durations are relatively randomly dispersed around a range of durations in an area limited by approximately two z-scores, reflecting a lack of structuring into larger units such as feet. English syllable durations are distributed in an L-shaped formation, with a much larger dispersion and a large cluster of relations between shorter neighbouring syllables in the bottom left quadrant, presumably correlating with sequences of unstressed syllables, as well as a fair number of long-short and short-long syllable pairs, indicating a higher level of structuring, e.g. into feet. There are very few long-long syllable pairs.

Summary table of global and accumulated TG duration functions (some do make sense...)
Time Group criterion: pausegroup, local threshold: 10, Min valid TG length: 2
Only inter-pause intervals measured; pauses not included

Overall duration:	**48504**	**Overall raw longer, ms:**	**15401**	**Overall raw shorter, ms:**	**14521**
Overall min:	**20.00**	**Overall max:**	**990.00**	**Overall range:**	**970.00**
Valid Time Groups:	**34**	**Overall rate/sec:**	**5.67**		
Components: global tendencies					
Overall mean:	**176.38**	**Overall median:**	**150.00**	**Overall SD:**	**113.58**
Overall npvi:	**62.00**	**Overall intercept:**	**156.12**	**Overall slope:**	**0.15**
Mean of means:	**182.18**	**Median of means:**	**176.70**	**SD of means:**	**34.75**
Mean of medians:	**168.68**	**Median of medians:**	**160.00**	**SD of medians:**	**40.88**
Mean of SDs:	**90.02**	**Median of SDs:**	**86.16**	**SD of SDs:**	**39.87**
Mean of nPVIs:	**60.00**	**Median of mnPVIs:**	**51.00**	**SD of nPVIs:**	**17.91**
Mean of intercepts:	**143.59**	**Median of intercepts:**	**130.80**	**SD of intercepts:**	**71.16**
Mean of slopes:	**10.65**	**Median of slopes:**	**11.86**	**SD of slopes:**	**41.10**
Components: correlations					
mean::TGdur:	**-0.190**	**median::TGdur:**	**-0.427**	**SD::TGdur:**	**0.230**
nPVI::TGdur:	**0.097**	**slope::TGdur:**	**0.061**	**intercept::TGdur:**	**-0.178**
nPVI::mean:	**0.128**	**slope::mean:**	**0.028**	**intercept::mean:**	**0.503**
nPVI::median:	**0.026**	**slope::median:**	**0.005**	**intercept::median:**	**0.310**
nPVI::SD:	**0.383**	**slope::SD:**	**0.051**	**intercept::SD:**	**0.229**

Figure 7: Screenshot of global statistics over a sequence of interpausal units.



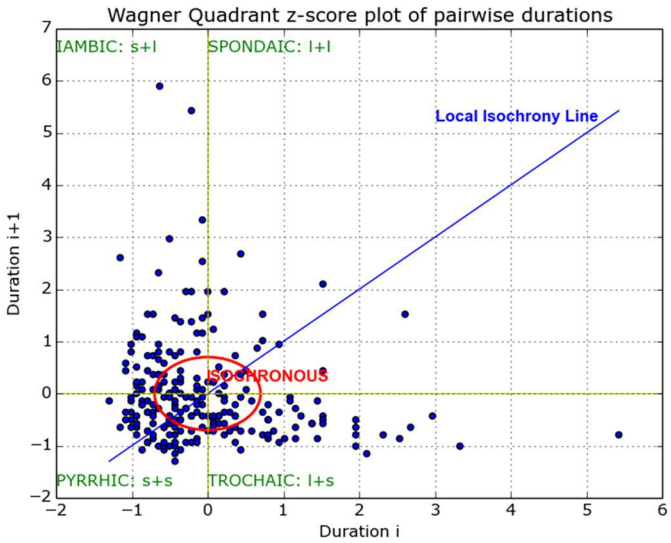


Figure 8: Wagner Quadrant Graphs for Mandarin and English syllable durations in similar reading genres.

3.3.2.6 Reformatted data and analysis outputs. A number of additional options are provided for converting the input data and calculated values (e.g. duration differences, z-scores, DDTs, statistics) into character separated value (CSV) formats, which are convenient for further processing with spreadsheets and other statistical tools. One of the CSV outputs, whether derived from a Praat or CSV input format, has a format identical to a CSV input format, tested by 'recycling' as input to the TGA, leading to identical outputs for all analyses.

3.4 Application in phonetic studies as TGA evaluation

3.4.1 Overview. The TGA online tool has been used in a number of published studies, which may count as a form of functional evaluation of the tool. The most interesting applications have been in studies of native and non-native varieties of Mandarin Chinese, but other applications have been made to genres in English, to Polish and to the Niger-Congo language Tem (ISO 639-3 kdh), a language of Togo (Klessa et al., 2014; Gibbon et al., 2014; Yu, 2013; Yu & Gibbon, 2012, Yu et al., 2014, 2015).

In the following subsections, two constrastive studies are outlined, on native vs. dialect-accented Mandarin, and on the proficiency levels of Mandarin L2 non-native vs. native L1 English pronunciation.

3.4.2 Dialect-accented Mandarin vs. Standard Beijing Mandarin. A pilot annotation mining experiment was undertaken with recordings of 6 speakers (3 from the Hangzhou area and 3 from Beijing) reading a Mandarin Chinese translation of the IPA standard text 'The North Wind and the Sun', taken from the CASS corpus.

Time Tree (TT) relations (Gibbon 2006) over interpausal groups were investigated. The following brief example shows a quasi-iambic TT (represented as bracketing) of

the Mandarin utterance “zhe4 shi1hou5, lu4 shang5 lai2 le5 ge4 zou3 daor4 de5” (at that time, on the street came a traveller), and a grammatical bracketing:

Quasi-iambic TT (the numbers represent tones):

(((zhe4 (shi2 hou5)) (((lu4 shang5) (lai2 (le5 (ge4 zou3)))) daor4)) (de5 PAUSE))

Grammatical bracketing:

((zhe (shi hou)), (lu shang) ((lai) (le) (ge) (zou daor de)))

A comparison of the TT bracketing and the grammatical bracketing (shi2 hou5) and (lu4 shang5) in the TT correspond to the words (shi hou) and (lu shang) in the grammatical bracketing.

Different trees were constructed based on different local thresholds for syllable duration differences, from 10ms to 220ms. Relations between the different trees and words of one or more characters/syllables were investigated. The percentage of agreement between tree constituents and words is shown in Figure 9 as a function of duration difference thresholds (DDTs), for three Hangzhou dialect speakers (HD) and three Mandarin (MD) speakers.

Below a duration difference threshold of about 50 ms, correspondences between syllable groups and words are low, and are comparable among speakers. Correspondences gradually increase and begin to diverge until about 100 ms, where they rapidly increase and interesting patterns emerge: (1) correspondences for Beijing Mandarin remain similar as thresholds move beyond 50 ms; (2) for the Hangzhou variety they are more diverse, as would be expected in a comparison between a standard accent (Beijing Mandarin) and a non-standard regional accent (Hangzhou Mandarin).

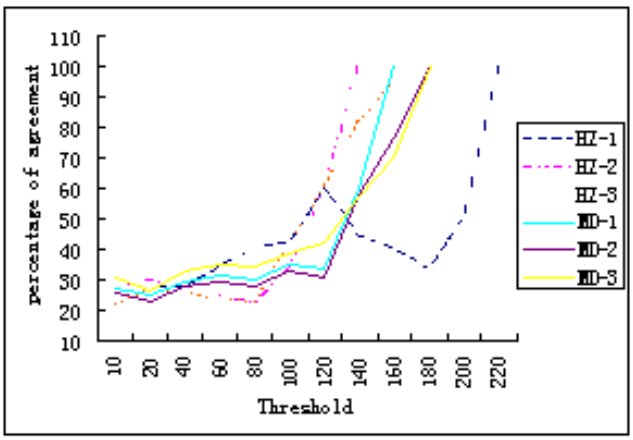


Figure 9: Relations between duration-based syllable groupings and words for speakers of Beijing and Hangzhou varieties of Mandarin Chinese.

3.4.3 Chinese EFL learners vs. English native speakers. Speech recordings of 20 Chinese L2 speakers and 10 English native speakers were used. First the

proficiency of the non-native speakers was graded by expert native and non-native English teachers into *poor*, *medium* and *advanced* groups. Using the TGA, the data time-stamps in the annotation files were then further investigated for temporal properties nPVI and syllable rate and temporal patterns. The results are shown in Table 4. The variability of both male and female Chinese learner groups are clearly functions of the proficiency level, while the proficiency of the female learners is somewhat higher by these measures.

Table 4: Summary of mean variability and mean syllable rate for female (F) and male (M) reader groups.

		F: *nPVI*	**F: syll rate**	**M: *nPVI***	**M: syll rate**
Ch L2	**poor**	56	4.2	59	4.3
	medium	62	4.7	65	4.9
	advanced	73	6.3	-	-
Eng	**native**	73	5.3	73	4.8

Wagner Quadrant graphs were constructed with the same data, and show interesting differences in the distribution of adjacent syllable durations (Figure 10). The important feature of the figures is the overall distribution shape, not the details. The low proficiency speaker shows a random distribution of values through the four quadrants. The English native speaker, on the other hand, tends to cluster values in the shorter-shorter, shorter-longer and longer-shorter quadrants; the overall pattern is L-shaped, with larger dispersion range. The advanced Chinese speaker also shows an approximate L-shaped distribution, but small dispersion range. The L-shaped distributions reflect anisochronous syllable timing in English, and the clustering in the shorter-shorter quadrant could be interpreted as sequences of unstressed syllables, indicating non-binary foot structures. Further research is needed to investigate this claim.

Additionally, duration difference token (DDT) *n*-grams were investigated. Percentages for purely alternating quadgrams and quingrams were calculated for each speaker (Table 5). The number of strict quadgram alternations appears as a function of proficiency. Quingrams show no obvious tendency. The non-natives have far fewer strictly alternating sequences than the English native speakers.

Finally, percentages of time-tree/grammar matching between Chinese L2 learners and native speakers were compared in respect of matching and proficiency. Results are shown in Table 6; matchings and proficiency correlate, $r^2 = 0.955$, $p < 0.01$.

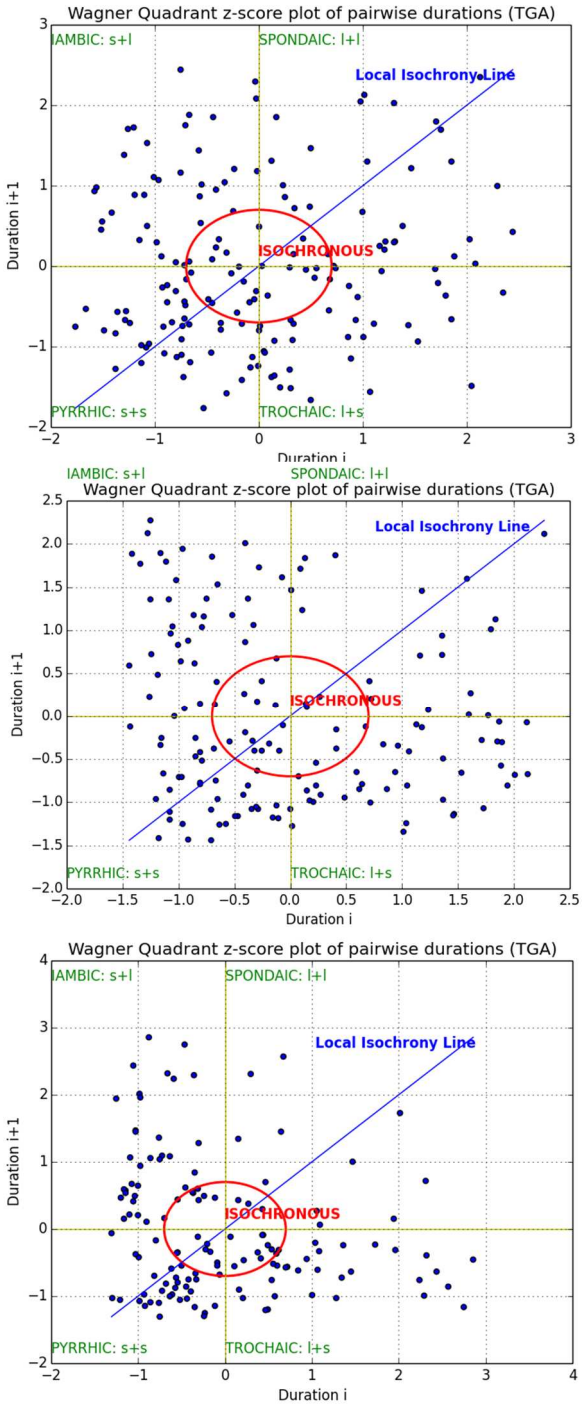


Figure 10: Automatically generated WQ graphs for Chinese L2 English, poor, female; Chinese L2 English, advanced, female; native speaker (USA), female (dispersion shapes are important in the figures, not details).

Table 5: Temporal quadgram and quingram alternation.

		F: 4-gram	F: 5-gram	M: 4-gram	M: 5-gram
Chinese	**poor**	4.5	1.7	5.1	02.6
	medium	8.5	4.3	2.3	08.5
	advanced	5.1	5.8	-	12.2
English	**native**	2.6	2.4	-	09.8

Table 6: Average Time Tree - grammar correspondences.

		female	male
Chinese	**poor**	65.80	67.08
Chinese	**medium**	72.40	69.20
Chinese	**advanced**	75.40	-
English	**native**	77.00	76.95

4 Summary and outlook

The present contribution provides an overview of relevant methodologies for analyzing temporal structures by means of annotation mining with annotated speech data leading to the specification, design, implementation and application of an online tool, Time Group Analyzer (TGA), for the support of linguistic phonetic analysis of speech timing, using time-stamped data, are described. The online tool provides extensive basic statistical information, including linear regression (for duration slope, i.e. acceleration and deceleration) and correlations between the different statistics over sets of Time Groups defined as interpausal units or dynamic (accelerating or decelerating) units. Three innovative visualizations are introduced: Δ*dur* duration difference tokens; top-suspension column charts for Δ*t* and Δ*dur* visualization, and Δ*dur* based *Time Trees* , which are represented as nested parentheses.

Informal evaluation of usability by four trained phoneticians and field evaluation is demonstrated by successful use in published studies, as well as adoption of modules of the TGA tool in software by other developers (AnnotationPro, SPPAS, this volume). The TGA tool reduces previous analysis times for mining time-stamped annotations by several orders of magnitude and supports the achievement of insightful results.

An offline version of TGA for processing large annotation corpora rather than single files is undergoing testing, and further functions such as box plots for timing distributions are in progress.

We anticipate further applications in the L2 teaching field for materials design and proficiency testing, and for the development of models for speech technology. Other disciplines which use duration metrics, such as forensic phonetics, clinical linguistics, dialectometry, stylometry and language acquisition, are also expected to benefit from the efficient methodology provided by the TGA.

References

Abercrombie, D. (1964). Syllable quality and enclitics in English. In: Abercrombie, D., Fry, D. B., McCarthy, P. A. D., Scott, N.C., & Trim, J.L.M. (Eds.): *In Honour of Daniel Jones.*

London: Longmans, 216-222.
Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University.
Arvaniti, A. 2009. Rhythm, Timing and the Timing of Rhythm. *Phonetica* 2009, 66, 46–63.
Auran, C., Bouzon, C., & Hirst, D. J. (2004). The Aix-MARSEC project: an evolutive database of spoken British English. *Speech Prosody* 2004, Nara, 561-564.
Barbosa, P. A. 2002. Explaining Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm production. In *Cadernos de Estudos Lingüísticos* 43, 71-92. Campinas.
Bird, S., Klein E., & Loper, E.. (2009). E. *Natural Language Processing with Python*. Beijing, etc.: O'Reilly.
Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.
Buschmeier, H., & Wlodarczak, M.. (2013). TextGridTools: A TextGrid Processing and Analysis Toolkit for Python. *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013),* Bielefeld, Germany, 152–15.
Carson-Berndsen, J. (1998). *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Dordrecht: Kluwer Academic Publishers.
Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics,* 37(1), 16-28.
Gibbon, D. (2003). Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of ICPhS 15, Barcelona, 2003*.
Gibbon, D. (2003). Corpus-based syntax-prosody tree matching. In *Proceedings of Eurospeech 2003, Geneva 2003*. 761-764.
Gibbon, D. (2006). Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In: Sudhoff, S., D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter and J. Schließer, Eds.*Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 281-209.
Gibbon, D. (2013). TGA: a web tool for Time Group Analysis. In Hirst, D. J., & Bigi, B. (Eds.) *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013*. 66-69.
Gibbon, D., & Gut, U. (2001). Measuring speech rhythm. *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 91-94.
Gibbon D., Hirst, D., & Campbell, N. (Eds.). (2012). *Rhythm, Melody and Harmony in Speech: Studies in Honour of Wiktor Jassem, Special edition of Speech and Language Technology* 14/15. Poznań: Polish Phonetics Society.
Gibbon, D., Klessa, K., & Bachan, J. (2014). Duration and speed in speech events. In: Mikołajczak- Matyja, N., & Karpiński, M., (Eds.). (2013). *Studies in Phonetics and Psycholinguistics. In honour of Prof. Piotra Łobacz.* Poznań: Adam Mickiewicz Press. 59-83.
Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Oxford: Basil Blackwell.
Gut, U. (2012). *Rhythm in L2 speech*. In Gibbon D., Hirst, D., & Campbell, N. (Eds.): *Rhythm, Melody and Harmony in Speech: Studies in Honour of Wiktor Jassem, Special edition of Speech and Language Technology* 14/15. Poznań: Polish Phonetics Society. 83-94.
Hirst, D. (2009). The rhythm of text and the rhythm of utterances: From metrics to models. *Proceedings of Interspeech 2009, Brighton*. 1519-1523.
Jassem, W. (1952). *Intonation of conversational English (Educated Southern British).*Wrocław: Wrocławskie Towarzystwo Naukowe.
Jassem, W., Hill, D. R., & Witten, I. H. (1984). Isochrony in English speech: Its statistical validity and linguistic relevance. In: Gibbon, D., & Richter, H. (Eds.): *Intonation, accent and rhythm: studies in discourse phonology*. Berlin: Walter de Gruyter, 203–225.
Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. In Miyake, N., Peebles, D. & Cooper, R. P. (Eds.): *Proceedings of the 34th Annual Conference of the Cognitive Science Society, Austin, TX*: Cognitive Science Society.

Klessa, K., & Gibbon, D. (2014). Annotation Pro + TGA: automation of speech timing analysis. *Proceedings of LREC 2014, Reykjavik*. Paris: ELDA.

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V., & Chen, X.. (2000). CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In *Proc. Interspeech 2000*, 485-488, Beijing.

Low, E. L., Grabe, E., & Nolan, F.. (2000). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. Language and Speech, 43(4), 377–401.

Pike, K. L. (1945). The Intonation of American English. Ann Arbor.

Ramus, F., Nespor, M., & Mehler, J.. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265-292.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In Crystal, D., (Ed.): *Linguistic Controversies: Essays in Linguistic Theory and Practice*. London: Edward Arnold, 73–79.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1986). On the measurement of rhythmic irregularity: a reply to Benguerel, *Journal of Phonetics*, 14, 327–330.

Wagner, P. (2007). Visualizing levels of rhythmic organisation. *Proceedings of the International Congress of Phonetic Sciences, Saarbrücken 2007*, 1113-1116.

Włodarczak, M., Simko, J., & Wagner, P. (2012). Temporal entrainment in overlapped speech: *Cross-linguistic study*. *Proceedings of Interspeech 2012*. 615-618.

Yu, J. (2013). Timing analysis with the help of SPPAS and TGA tools, *Proceedings of TRASP 2013, Aix-en-Provence*. 70-73.

Yu, J., & Gibbon, D.. (2012). Criteria for database and tool design for speech timing analysis with special reference to Mandarin. *Proceedings of Oriental COCOSDA 2012, Macau* (IEEExplore Conf ID 21048). 41-46.

Yu, J., Gibbon D., & Klessa, K.. (2014). Computational annotation-mining of syllable durations in speech varieties. *Proceedings of 7th Speech Prosody Conference, 20-23 May 2014. Dublin*. 443-447.

Yu, J., & Gibbon, D. (2015). How natural is Chinese L2 English prosody? *Proceedings of ICPhS 2015, Glasgow*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0304.pdf

COMPARING PITCH DISTRIBUTIONS USING PRAAT AND R

Mietta Lennes[1], Melisa Stevanovic[2], Daniel Aalto[3], and Pertti Palo[4]
[1] Department of Modern Languages, University of Helsinki, Finland
[2] Finnish Centre of Excellence on Intersubjectivity in Interaction, University of Helsinki, Finland
[3] Rehabilitation Medicine, Communication Sciences and Disorders, University of Alberta, and Institute for reconstructive sciences in medicine, Misericordia Community Hospital, Edmonton, Canada
[4] Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, United Kingdom
e-mail: mietta.lennes@helsinki.fi, melisa.stevanovic@helsinki.fi, aalto@ualberta.ca, pertti.palo@gmail.com

Abstract

Pitch analysis tools are used widely in order to measure and to visualize the melodic aspects of speech. The resulting pitch contours can serve various research interests linked with speech prosody, such as intonational phonology, interaction in conversation, emotion analysis, language learning and singing. Due to physiological differences and individual habits, speakers tend to differ in their typical pitch ranges. As a consequence, pitch analysis results are not always easy to interpret and to compare among speakers.

In this study, we use the Praat program (Boersma & Weenink 2015) for analyzing pitch in samples of conversational Finnish speech and we use the R statistical programming environment (R Core Team, 2014) for further analysis and visualization. We first describe the general shapes of the speaker-specific pitch distributions and see whether and how the distributions vary between individuals. A bootstrapping method is applied to discover the minimal amount of speech that is necessary in order to reliably determine the pitch mean, median and mode for an individual speaker. The scripts and code written for the Praat program and for the R statistical programming environment are made available under an open license for experimenting with other speech samples. The datasets produced with the Praat script will also be made available for further studies.

1 Introduction

The analysis of the melodic aspects of speech serves various research interests, such as intonational phonology, speech communication, interactional linguistics, interactional sociology, emotion analysis and language learning. Relative pitch levels and patterns can be connected with many language-specific linguistic functions, such as intonation, stress, (sentence) accent or lexical tones. In conversation, subtle

variations in pitch have been shown to convey, for example, turn-taking or turn-yielding intentions (Duncan, 1972; Ford & Thompson, 1996; Szczepek-Reed, 2004), sequence organization (Kaimaki, 2010; Persson 2013), information status (Breen et al., 2010) and confidence (Scherer et al., 1973).

The pitch range of a speaker depends on physiological (Titze, 1989) and psycho-social (e.g., Cartei et al., 2014; Munson et al., 2015) factors and can serve as an identifying characteristic of the speaker (Kinoshita et al., 2009; Munson, 2007). Due to this variability, theories of intonational phonology usually work with relative pitch levels or excursions within utterances (see, e.g., Ladd, 1996 for a detailed discussion) and not absolute pitch. Moreover, the functional significance of pitch in conversation depends not only on its absolute levels but largely on its relation to the speaker-specific pitch range (e.g., Couper-Kuhlen, 1996). In other words, what counts as high or low varies by speaker (Leather, 1983; Moore & Jongman, 1997). These insights are supported by empirical research showing that listeners are capable of locating the pitch of a given speech sound within the speaker's range without external context or previous exposure to the speaker's voice (Honorof & Whalen, 2005). Thus, in order to analyze the pitch of a given speaker, it is necessary to relate it to his or her typical pitch range.

Since the present study deals with perceptual and relative properties in speech, we prefer to use the term *pitch* instead of the acoustic concept of fundamental frequency (f_0) in this work. The choice of scale plays an important role in analyzing pitch variation. Fundamental frequency f_0, which correlates non-linearly with the perceived pitch in voiced speech, is measured and reported as absolute values in Hertz scale. Traunmüller and Eriksson (1995) provide an overview of previous reports concerning the f_0 ranges of male and female speakers. They point out that when the f_0 range is expressed in the absolute Hertz scale, female speakers appear to exhibit a wider range than men, but the difference more or less disappears when the data are converted into semitones. When expressed in semitone scale, the overall shapes of pitch distributions appear to be similar between speakers (Lennes, 2007) and even between different languages (Lennes et al., 2008). This is not surprising, since humans have similar vocal organs, and the vocal folds can only be stretched within certain limits. Moreover, during modal phonation, it is not possible to instantaneously jump from low pitch to high pitch or vice versa, but the speaker will have to glide through the intermediate pitch levels.

The aim of the present work is to investigate the general distribution of pitch in conversational Finnish speech and to discover the minimum requirements for obtaining reliable statistics of speaker-specific pitch ranges. We will first calculate and describe the pitch distributions of 40 Finnish speakers in everyday conversation, pinpointing some factors that may affect the typical distribution shape in individual cases. Using a bootstrapping method, we will then attempt to determine the minimum amount of samples that is required in order to calculate the mean, median or mode.

We invite other researchers to replicate the results and to extend and improve the method. For these purposes, the code for Praat and R, as well as the pitch data produced for this study, will be shared online under an open license. Our actual workflow is described more explicitly in the documentation of the scripts. Since the

tools may be of interest to readers without a background in phonetics, we will first briefly describe how human speakers may vary in their preferred pitch ranges and how automatic pitch analysis generally works.

2 Background

Speakers tend to differ in the pitch region they usually employ during speech. This variability in preferred pitch is partly due to anatomical and physiological differences. On average, men have longer and thicker vocal folds than women (e.g., Titze, 1989). This is largely why female speakers tend to speak at a higher pitch than male speakers. Similarly, small children tend to use a much higher pitch region than adults.

In addition to the aforementioned physical differences, people also exhibit culture-specific and idiosyncratic ways of using their voice while speaking or singing. Some speakers may be perceived to have "lively" voices, whereas others may sound "monotonous". This may mean that some speakers employ larger pitch ranges, whereas others prefer to keep their pitch close to their personal level of comfort. On the other hand, some speakers creak almost all the time, whereas others use a breathy voice quality or one that may sound like falsetto. In various medical conditions or as a consequence of a surgical treatment affecting the upper airways, the pitch of a person's voice may change significantly. All in all, voice and pitch are an important part of a person's self and identity.

Since people are apparently able to estimate the general height of each others' voices almost instantly, it is likely that this impression is not based on, e.g., the highest and lowest pitches, which would vary from one utterance and situation to the next. Instead, listeners are more likely to "tune in" to the pitch region that the speaker uses most of the time. In music, the typical, most comfortable pitch range of a singer is sometimes referred to as the *tessitura*.

Thus, in order to be able to compare speakers reliably, it is necessary to determine the typical or preferred pitch range of a particular speaker. However, this is not a technically straightforward task. The automatic analysis of pitch or fundamental frequency in speech does not always provide data that can be easily interpreted and compared among speakers. In addition, poor technical quality of the speech material can distort the analysis result. In order to get plausible data, researchers need to be aware of the general properties and inherent limitations of the pitch extraction algorithm that is being applied.

2.1 Automatic pitch detection

In automatic pitch analysis, the voiced portions of speech are expected to represent a single quasi-periodic sound source. This is true in recordings where only one speaker is speaking at a time and the speaker's vocal folds are vibrating normally and rather steadily. Pitch analysis is usually tuned so as to pick up the fundamental frequency f_0, which usually corresponds to detecting the presence and frequency of the slowest, at least nearly periodic component in the complex acoustic signal. At least during modal (regular) phonation, the f_0 thus reflects the frequency of the glottal pulses, i.e., the repetitive opening-closing sequences of the vocal folds.

There are various methods available for automatic pitch extraction and for representing the resulting pitch contours. In this study, we apply the standard autocorrelation method available in the Praat program (the command **To Pitch...**). This method is often used for studying intonation in speech, whereas the cross-correlation algorithm, also available in Praat, is suited for special purposes, such as voice analysis. In practice, both algorithms calculate a sequence of pitch values using short, partly overlapping time windows or frames extracted from the original audio signal. The resulting values can be plotted as a pitch contour as a function of time, or they may be further analyzed.

Since the larynx and the articulatory organs are rarely held completely steady during speech, the frequency structure of the speech signal changes practically all the time. Each analysis window may include speech that is only partly voiced and/or where the f_0 is changing. In order to be able to select the best or most plausible candidate among a number of all possible pitch candidates within each analysis window, the pitch algorithm requires the user to supply the minimum and maximum frequencies prior to the analysis. These parameters can be adjusted according to the expected frequencies for a particular speaker or for specific analysis purposes. The minimum frequency parameter defines the duration of each analysis window. In order to detect a low f_0, where the glottal periods are relatively long, the analysis window needs to be wider than for a high f_0. However, in case the minimum parameter is set too low, the wide analysis frame will conceal fast changes in the f_0. In addition, users can also adjust more advanced parameters that control the tolerance for abrupt pitch changes between consecutive analysis frames. These parameters are used in the pitch algorithm, since human speakers are not able to shift the pitch of their voice up or down at an arbitrary rate. Nevertheless, it is to be noted that even if all the parameters are set in an appropriate way, external noises and overlapping speakers may distort the result.

Non-modal phonation, such as creaky voice, occurs quite frequently in everyday talk (Ogden, 2001; Gobl & Ní Chasaide, 2003; Yuasa, 2010). Irregular periodicity or two simultaneous glottal modes of vibration may occur during creaky or glottalized phonation, and they are difficult to analyze consistently with the standard pitch algorithms. Such events will often result in missing values, potentially erroneous values with halved or doubled frequency (often referred to as “octave jumps”), or other outliers in the pitch curve. In these cases, it is still possible to perform a partly manual analysis in Praat in order to check the result. This can be accomplished for instance by editing a Pitch object. Alternatively, a PointProcess object can first be generated from the Pitch and the corresponding Sound object. Next, the locations of the automatically detected pitch periods can be edited in the PointProcess editor, after which the PointProcess can be converted back to a Pitch object. Manual editing is applied for instance in the ProsodyPro system, which is intended for the analysis of pitch contours on more large-scale material (Xu, 2013). However, manual work is time-consuming, somewhat subjective, and error-prone. On the other hand, it would be efficient to analyze large amounts of data in batch mode, but even if the pitch analysis parameters are individually adjusted for each speaker, it may not be

ultimately possible to avoid the halved or doubled frequency values. It would be useful to be able to automatically discover which regions of the pitch distribution are likely to represent the speaker's modal voice and which parts are potentially less reliable.

3 Material

We built our analysis on two corpora of conversational speech. The FinDialogue corpus, a part of the larger FinINTAS corpus, contains ten conversations (five male-male dyads and five female-female dyads) between young, native Finnish-speaking adults. The participants in each dialogue knew each other well. The dialogues were recorded in an anechoic room using high-quality headset microphones (AKG HSC-200 SR). The two speakers in each dialogue were sitting a few meters apart and facing opposite directions. They were instructed to chat freely for 45-55 minutes either on a few given topics or on whatever they felt like talking about. Each speaker's voice was recorded with a DAT recorder (Tascam DA-P1) on a separate track in a stereo file and downsampled to a rate of 22050 Hz (sample size 16 bit). Thus, it was possible to analyze each speaker's voice in isolation when required. This corpus will be referred to as Corpus A.

The other collection of conversational Finnish speech, which we shall call Corpus B, consists of shorter dialogues with 8 adult female and 8 adult male speakers (3 male-male dyads, 3 female-female dyads, and 2 male-female dyads). The dialogues were recorded in various conditions using one or two microphones. The dialogues included two mundane telephone conversations (2-3 minutes each), two informal planning interactions in a workplace setting (5 minutes and 20 minutes), and three conversations, where the participants were engaged in a joint decision-making task in an experimental setting (2-4 minutes each). The speakers are referred to with a number preceded by the letter F for female and M for male speakers.

4 Analysis

The analysis procedure of this study was implemented as two main scripts: one for collecting the pitch data from the original audio files in Praat, and the other for running various analyses on the pitch data and for plotting the figures using the R statistical programming environment. The two scripts are available and documented on GitHub (Lennes, 2016).

Using a Praat script (see Lennes, 2016 for a detailed description), all the audio files were analyzed with the standard, autocorrelation-based pitch algorithm available in Praat. The distance between consecutive analysis frames was set to 0.02 seconds, resulting in 50 observed pitch values per second in the measured data.

In a first analysis pass, the default minimum frequency parameter was set at 50 Hz and the maximum at 600 Hz. (The default parameters can be changed in the Praat script for other experiments.) These parameters would be too far apart for almost all adult speakers, i.e., the minimum would be clearly below the lowest fundamental frequency that most male speakers would tend to use, and the maximum value would exceed most of the f_0 values of female speakers. The intention was that these settings

would be likely to create anomalies in the initial pitch data. After this first analysis pass, speaker-specific minimum and maximum frequencies were manually determined by inspecting the pitch distributions in R and by locating and generously delineating the pitch cluster with the highest density in each distribution. The speaker-specific parameters were applied in the second analysis pass so as not to include extremely low or high pitch values.

In total, three different datasets were obtained. Dataset 1 was calculated from raw audio using the default minimum and maximum parameters. This type of analysis can, in principle, be done for any audio file without knowing anything of the speaker(s), although the results will not be reliable. Dataset 2 was produced by applying the speaker-specific pitch parameters to analyze the raw audio. This way, it was possible to see how the pitch distribution was affected by whether the minimum and maximum parameters were set individually or not. In order to save some disk space, all undefined pitch values were excluded from these first two datasets. It should be noted that Datasets 1 and 2 are considered as experimental and they will not be useful for audio files that include more than one speaker. Dataset 3 was calculated from the annotated corpora so that only those parts of the audio signals were analyzed where the speaker in question was actually speaking, according to the utterance-level annotations in the TextGrid files. Dataset 3 was used for comparing speaker-specific distributions.

The frequency values from all the individual analysis frames obtained for all three datasets were automatically written to data tables (tabulated text files) in both Hertz and semitones with respect to the frequency of 100 Hz. For Dataset 3, a total of 489,485 pitch analysis frames, including 277,384 voiced ones, were recorded. A pitch difference expressed in semitones corresponds to the respective musical interval, which makes the data easier to read and interpret. For instance, a difference of 12 semitones (ST) corresponds to an octave, an interval of 7 ST corresponds to a perfect fifth and 5 ST to a perfect fourth. In this article, all pitch values expressed in semitones are provided relative to 100 Hz, unless another reference level or comparison is mentioned.

5 Results

The analysis continued by visualizing the general properties of the pitch distributions for each individual speaker. Since our aim was to estimate the shape of the overall pitch distributions of individual speakers and since pitch and frequency are continuous variables, we first plotted the probability density curves for all speakers and for all three datasets for inspection. A density plot is a continuous version of the more familiar histogram.

5.1 Probability density

The pitch distributions for one female speaker (F3 in Corpus A) are plotted in Figure 1. The analysis calculated from the unannotated audio (Dataset 1) is indicated with a dotted line, Dataset 2 with a dashed line, and Dataset 3 with a solid line. It is observed that the main distribution is skewed to the right. The speaker generally stays around her typical pitch level (mode = 9.8 ST, 176 Hz), but she sometimes goes approximately 6 semitones below or 12 semitones above her mode. Since the audio

signal was of high technical quality, this is probably why there is very little difference between the Dataset 1 distribution, calculated from raw audio with the default parameters, and the result of the more speaker-specific analysis in Dataset 3.

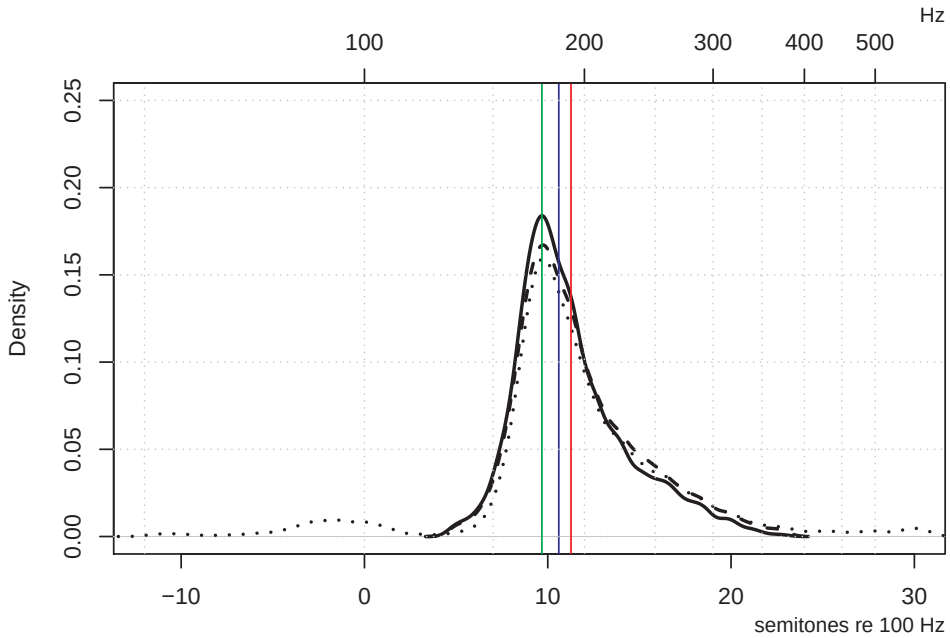


Figure 1: The probability density function of the pitch values obtained from conversational speech recorded from one female speaker (F3 in Corpus A). The mean pitch (11.27 semitones above 100 Hz) is indicated with a red vertical line, median (10.6 ST) with blue and the pitch mode (9.68 ST) with a green line. The corresponding values in the absolute Hertz scale are 195 Hz, 184 Hz and 174 Hz. The dotted line represents the pitch distribution obtained from raw audio (Dataset 1), the dashed line is the distribution calculated from raw audio with manually defined speaker-specific parameters (Dataset 2), and the solid line represents the data calculated within annotated utterances only (Dataset 3).

Another example of the pitch distributions is shown in Figure 2 for the female speaker F23 in Corpus B. In this case, Dataset 1 includes an external low-frequency noise. The total amount of data for this speaker was small (2282 samples in Dataset 3), which is probably the reason why the distribution looks more irregular than that of speaker F3 (12640 samples).

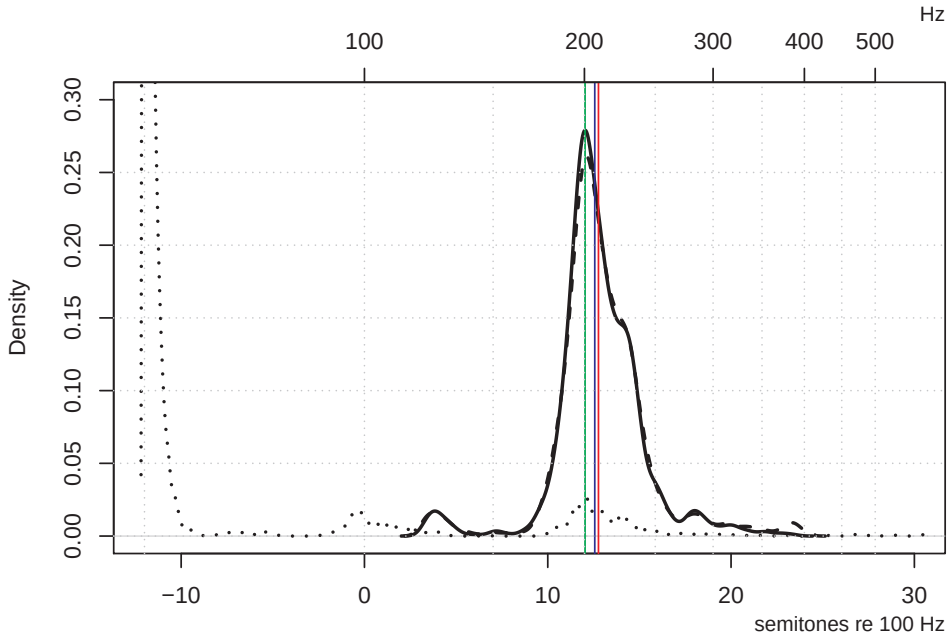


Figure 2: The pitch distribution of the speaker F23 (Corpus B), whose recording contained a constant, low, humming background noise at around 50 Hz. The effect of the noise is prominent in the raw overall pitch distribution (Dataset 1, dotted line), where the minimum frequency parameter was set at 50 Hz.

The number of pitch frames analyzed for each speaker is provided in Table 1. A summary of their individual pitch statistics in Dataset 3 is provided in Table 2. As a general observation, it is seen that the pitch mode for the maximal data in Dataset 3 is in most cases (for 33 speakers out of 40) located below the median, which in turn is usually below the mean pitch for each speaker. This confirms that a majority of the distributions are skewed to the right. Only seven speakers (F1, F21, M4, M5, M21, M22 and M26) are different in this respect. M4 has an almost symmetric distribution, and M5 is even slightly skewed to the left. Both of them creaked quite extensively. M21 and M22 exhibit bimodal pitch distributions, which may be due to the technical quality of the audio, perhaps overlapping speech. M26 has a relatively flat and irregular distribution.

Table 1: The number of pitch frames recorded for each of the 40 speakers in Dataset 3. The speakers of Corpus A are shown in the two leftmost columns, and speakers in Corpus B in the rightmost ones.

Speaker	N	Speaker	N	Speaker	N	Speaker	N
0F1	16335	0M1	13387	F21	17296	M21	10139
0F2	15669	0M2	13409	F22	04757	M22	07712
0F3	12640	0M3	12144	F23	02282	M23	04900
0F4	15455	0M4	21201	F24	07829	M24	02790
0F5	14563	0M5	15405	F25	08424	M25	02729
0F6	19197	0M6	14313	F26	08846	M26	05053
0F7	15506	0M7	19685	F27	12056	M27	02397
0F8	15615	0M8	16024	F28	17702	M28	08638
0F9	15778	0M9	21053				
F10	09725	M10	10216				
F11	15921	M11	19740				
F12	14737	M12	08217				

5.2 Establishing a reference pitch for comparing speakers

Figure 3 shows the pitch densities of all 40 speakers in Dataset 3. It is observed that male speakers tend to have lower pitch than females, which is hardly surprising. The overall mean pitch in Dataset 3 was 191.3 Hz (10.8 ST) for female speakers and 117.5 (2.2 ST) for males, with all speakers pooled. The corresponding standard deviations were 44.6 Hz (3.8 ST) for females and 33.0 Hz (4.5 ST) for males. The pitch distributions for the individual males form a cluster around 100 Hz or below, and most of the distributions for females are centered at about 150-200 Hz. However, there are also 6 male and 2 female speakers with more clearly overlapping distributions whose modes are located between 100 and 150 Hz. It is thus not uncommon for the two genders to exhibit similar pitch. Another important observation is that the shapes of these primary distributions exhibit at least roughly similar properties: usually one peak, generally similar width, and the distributions are more or less right-skewed.

In order to compare the way different speakers exploit their typical pitch range, it is possible to shift the pitch distributions over each other by referring the semitone-scaled pitch values to the speaker-specific modes. Using the semitone scale and the mode as the common anchor point enables us to compare the details of the individual distributions, while no information is lost about the perceptual distances of the pitch values. The result is shown in Figure 4.

Figure 5 shows a histogram of the mode-referred pitch values pooled for all 40 speakers in Dataset 3, supplemented with the corresponding probability density curve. The pooled mean of mode-referred pitch was 1.14 ST (s = 3.36 ST, median 0.61 ST). In the histogram of the pooled data, the probability of the bin with the highest probability (-0.5–0.5 ST) was 0.17 (17 %). The sum of the probabilities of the bins between -2.5 ST and 4.5 ST was approximately 0.77.

Table 2: Summary statistics of the primary pitch distributions for 40 speakers (Sp.) in Dataset 3.

Sp.	Mode		Median		Mean		Stdev	
	ST	Hz	ST	Hz	ST	Hz	ST	Hz
F1	10.29	165.84	10.23	180.59	10.55	186.35	2.73	31.85
F2	09.47	172.06	09.64	174.49	10.00	180.35	2.63	29.69
F3	09.68	174.38	10.60	184.48	11.27	194.85	3.03	37.30
F4	11.29	189.61	12.86	210.23	13.36	221.05	3.50	48.64
F5	12.03	198.85	12.94	211.21	13.21	216.45	2.30	30.51
F6	09.59	172.85	10.47	183.08	11.12	194.11	3.39	43.01
F7	09.97	177.08	10.64	184.91	11.27	194.34	2.72	35.74
F8	09.69	173.93	11.24	191.46	11.81	202.78	3.73	48.71
F9	07.89	156.78	08.88	167.01	9.48	175.96	3.14	35.92
F10	03.06	118.57	03.61	123.19	3.90	127.05	2.85	22.83
F11	06.05	137.95	06.31	144.02	6.65	148.53	2.60	23.83
F12	07.06	150.13	07.42	153.51	7.65	156.72	2.01	20.19
F21	13.79	192.77	13.09	212.94	13.24	217.74	2.75	37.12
F22	10.73	184.95	11.28	191.85	11.78	201.13	3.25	40.83
F23	12.03	200.33	12.56	206.63	12.77	211.2	2.49	30.66
F24	10.44	182.09	11.19	190.84	11.91	202.25	2.99	40.26
F25	09.95	176.42	11.75	197.12	12.56	213.03	4.19	56.82
F26	13.04	210.59	13.59	219.23	13.95	229.00	3.63	51.17
F27	08.79	161.00	09.56	173.74	10.15	183.16	3.25	38.96
F28	09.00	166.97	10.05	178.72	10.71	189.75	3.53	41.96
M1	-1.22	093.09	-0.07	099.61	0.60	106.08	3.60	27.57
M2	-0.50	096.65	00.57	103.34	1.15	108.60	3.05	21.10
M3	-0.55	096.37	00.73	104.33	1.65	112.67	3.66	26.69
M4	06.80	147.51	06.79	147.98	7.06	152.78	3.03	29.65
M5	-0.49	095.56	-0.85	095.19	-0.32	104.91	5.49	52.94
M6	01.76	110.22	02.76	117.27	3.31	123.75	3.49	28.50
M7	-1.45	090.50	00.08	100.44	0.77	107.24	3.75	27.06
M8	-6.01	069.79	-5.93	071.00	-5.54	074.23	3.49	17.14
M9	04.45	128.15	04.81	132.05	5.16	136.33	2.65	21.79
M10	-3.07	083.35	-1.40	092.23	-0.74	097.53	3.17	20.10
M11	-0.04	098.49	01.42	108.52	2.16	116.59	4.01	29.92
M12	04.64	130.14	05.23	135.23	5.64	140.03	2.49	22.31
M21	02.08	112.21	01.56	109.40	1.43	111.37	3.85	25.48
M22	04.49	112.49	04.34	128.51	4.44	131.28	3.04	23.74
M23	01.26	106.74	02.43	115.07	2.86	119.52	2.77	20.52
M24	-1.03	093.87	00.14	100.81	1.18	109.72	3.71	26.46
M25	-0.48	096.75	00.63	103.73	1.45	110.72	3.19	22.71
M26	03.56	110.54	02.59	116.17	2.56	118.09	3.32	23.20
M27	-1.21	092.78	-0.52	097.06	0.05	102.22	3.28	21.30
M28	01.94	111.28	02.81	117.63	3.21	122.50	3.18	23.55

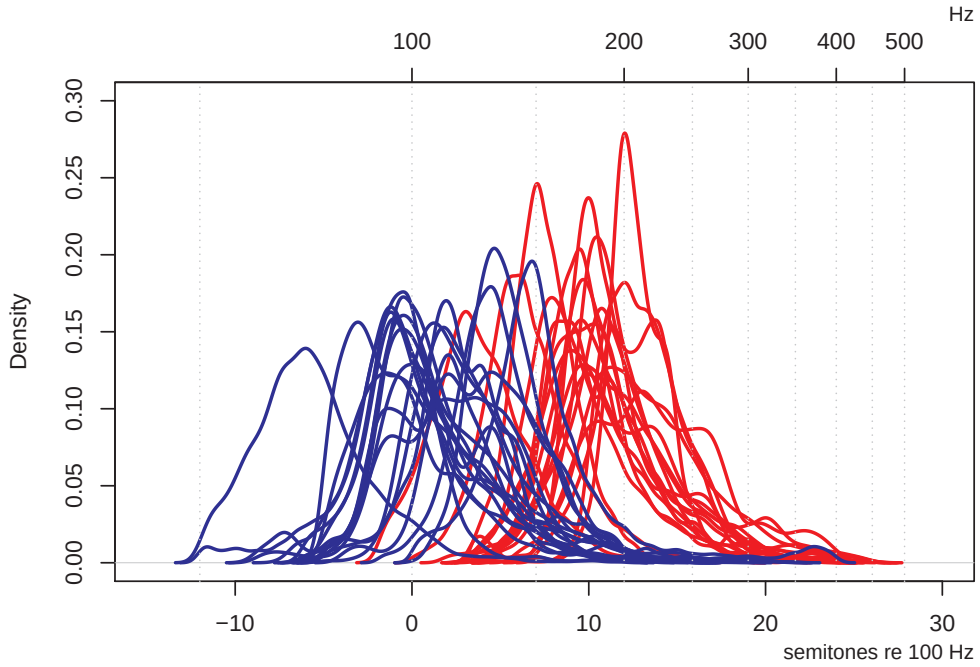


Figure 3: The overall pitch densities within annotated utterances of 20 male (blue lines) and 20 female (red) speakers.

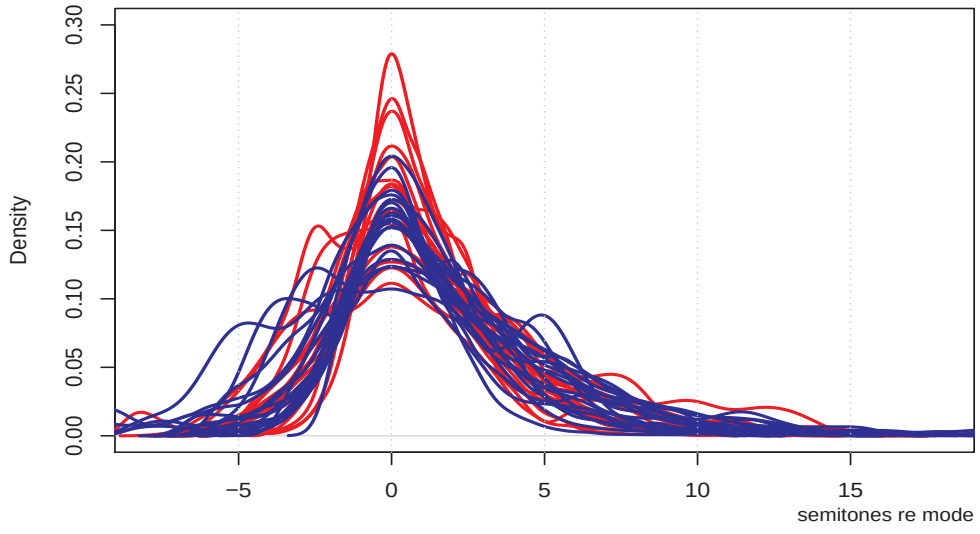


Figure 4: The mode-referred pitch distributions plotted as density curves for 40 speakers in Dataset 3. The zero pitch level refers to the speaker-specific mode. Male speakers are indicated with blue lines, females with red.

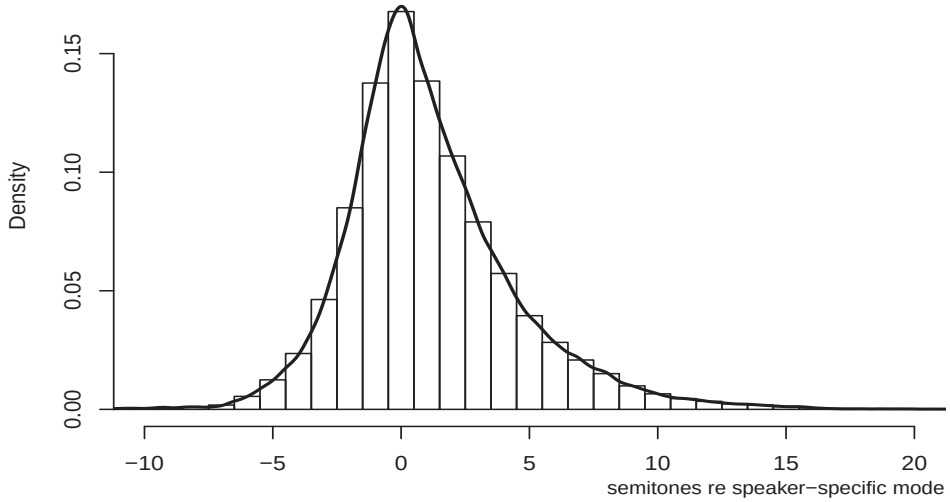


Figure 5: The distribution of mode-referred pitch values in voiced speech (N = 277,384) for all 40 speakers in Dataset 3. The zero pitch level refers to the speaker-specific mode. The bin width in the histogram is 1 semitone.

Thus, speakers would tend to exhibit pitch levels within such a span around their most typical pitch in about 77% of their voiced speech. 95% of all pitch values in Dataset 3 fall in the bins whose midpoints are located between -4 ST and 8 ST. Conversely, speakers would hit pitch levels outside this span in about 5% of their speech produced in the modal register. Since these probabilities are based on pooled data, they are to be taken as rough approximations. Speakers may differ to some extent, e.g., in the effective width of the primary pitch distribution.

5.3 Technical observations

Pitch analysis provides inconsistent results in cases where several speakers are captured in the same single-channel sound signal and two or more of them are speaking simultaneously. The analysis for the present study did not exclude the overlapped portions, since the amount of audible "crosstalk" in these dialogue corpora was considered relatively small and it only concerned a few speakers. However, such an exclusive feature could easily be implemented in the Praat script, when it is known which annotation tiers contain the utterance items that should not overlap.

The audio signal may sometimes contain background noise or electrical disturbances that can distort the pitch detection. For instance, in two of the dialogues in Corpus B, a humming noise was detected at the frequency of 50 Hz. This persistent noise is included in the analysis of Dataset 1 and thus creates an extra peak in the pitch distribution (see Figure 2 for an example). Since this kind of noise occurs within a low frequency range and usually does not overlap with speech frequencies, it might be possible to filter the noise out without significantly affecting the actual speech signal.

5.4 Bootstrapping

In order to estimate the minimum amount of speech that is required in order to obtain a reliable statistical description of the speaker's typical pitch range, we applied a bootstrapping procedure. In statistics, bootstrapping refers to any method – usually a statistic or a test – that uses random resampling of existing data. Bootstrapping can be used for calculating accuracy estimates of a (likewise estimated) statistic (Efron, 2003; Efron and Tibshirani, 1994). As such, it is used in finding sample sizes required for the convergence of a given statistical estimate that originates from an unknown distribution. In practice, random samples are drawn from a larger body of data. These samples are then analyzed as if they were regular samples from the studied phenomenon. For instance, it is possible to systematically increase sample size and repeat the random sampling a number of times for each sample size, and for each of these simulated samples to calculate the mean. This would provide a bootstrap estimate of the variation of the mean as a function of sample size and give us a way of estimating the sample size corresponding to a required level of accuracy.

For each of the 40 speakers, subsets of consecutive pitch values were randomly drawn from Dataset 3, beginning with the sample size of 50 pitch values (corresponding to 1 second of net speaking time) and increasing the sample size in steps of 50 values after each sampling round, either until the speaker had fewer samples than 1.5 times the sample size or until the maximum sample size of 10,000 pitch values was reached. For each sample size and for each speaker, up to five non-overlapping sequences of pitch values were drawn from the dataset, depending on whether a sufficient number of frames were available for the speaker in question. One single draw in the maximum sample size was possible for 16 speakers, who were represented with more than 15,000 pitch frames. The means of all the sampled portions from all 40 speakers are plotted in Figure 6, and the corresponding modes are shown in Figure 7. At sample sizes larger than 3000, fewer than five draws were possible for most speakers. However, the mean and mode have mostly converged before this point.

As shown in Figure 6, the standard deviation of the pitch means is about 2 ST in small sample sizes, but is reduced into less than 1 semitone after analyzing 650 pitch frames (only 12 seconds) or more. For many speakers, the pitch mode also converges quickly to a rather stable level and the overall standard deviation drops under 1 semitone after analyzing at least 34 seconds of net speaking time. For some speakers in Corpus B, the overall pitch distribution was bimodal, and the location of the primary mode is unstable, even after three minutes of net speaking time (e.g., speakers M21, M22, F21, F27). This phenomenon is visible in the mode-referred distributions that would overlap to a large extent apart from three female and two male speakers (see Figure 7). The bimodal distributions might be partly explained by the type of audio material. The recordings of the dialogues among M21 and F21, as well as F27 and F28, were noisy, the dialogues were recorded with only one microphone, and the speakers often overlapped in the signal. The reason for obtaining a bimodal pitch distribution in M22's recording was less clear, although background noise was present.

In some cases of Corpus B, the small amount of material available may explain why the distributions look unstable (cf. Table 1).

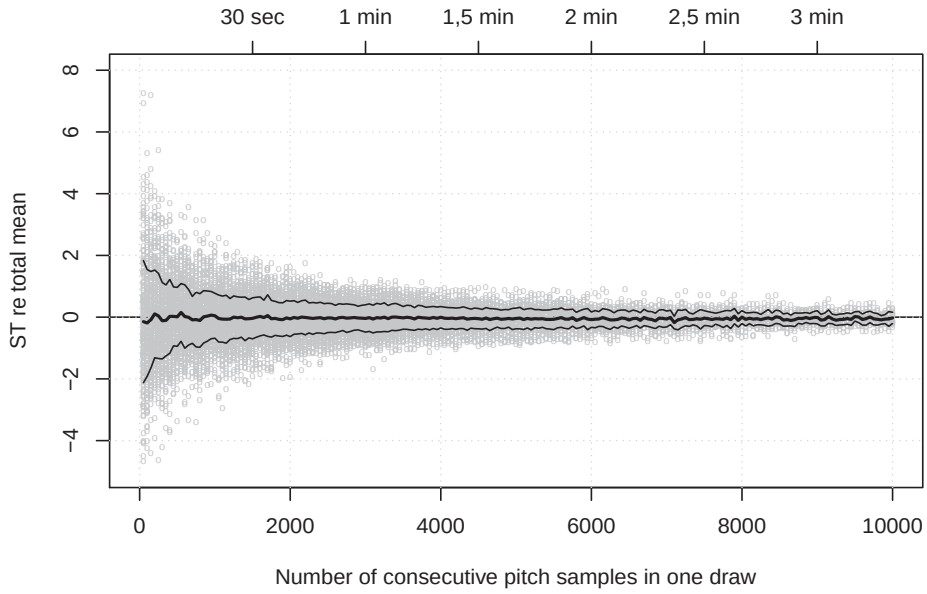


Figure 6: Bootstrapping the pitch mean. At most five sequences of 50 to 10000 consecutive pitch values were randomly drawn from each of the 40 speakers in Dataset 3. The **means** of all draws are plotted as grey circles relative to the corresponding speaker's total mean. The thick curve shows the local mean and the thin curves show the standard deviation for the means in each sample size. The values converge towards the speaker-specific mean in the complete dataset (zero level).

Figures 8 and 9 show the more detailed density curves in three exemplary conditions where each speaker is represented by one random sample of either 1000, 3000 or 6000 consecutive pitch points. In Figure 8, these pitch values are shown with respect to each speaker's overall pitch mean, and Figure 9 shows the corresponding mode-referred distributions. The mean of the pitch modes for the complete 1000-point samples was 0.12 ST (standard deviation 1.14 ST), 0.05 ST (s = 0.80 ST) for 3000 points and -0.11 ST (s = 0.43 ST) for 6000 points. The corresponding mean of the pitch means was 0.04 ST (s = 0.73 ST) for 1000, 0.04 ST (s = 0.44 ST) for 3000 and -0.08 ST (s = 0.25) for 6000 points.

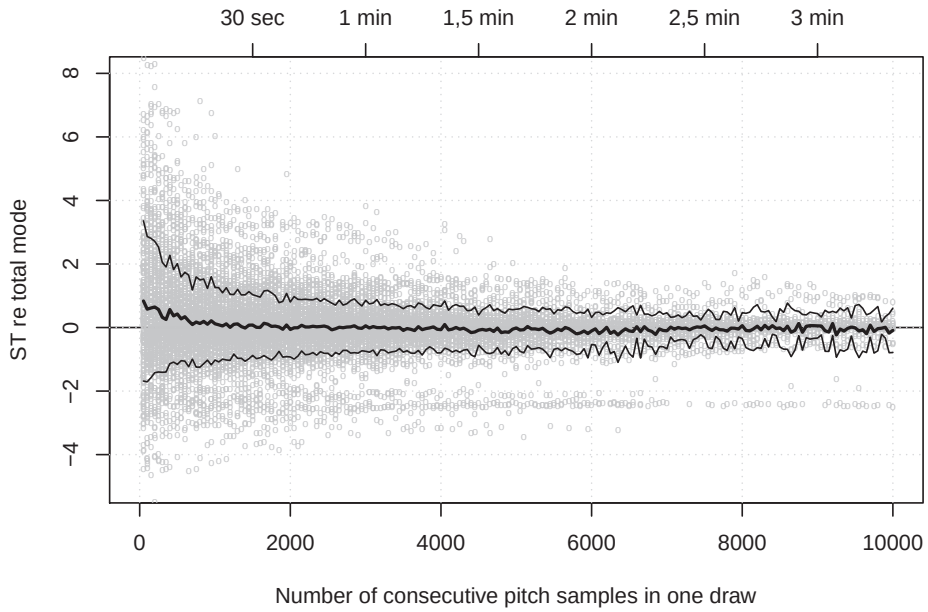


Figure 7: Bootstrapping the pitch mode. At most five sequences of 50 to 10000 consecutive pitch values were randomly drawn from each of the 40 speakers in Dataset 3. The **modes** of all draws are plotted as grey circles relative to the corresponding speaker's pitch mode in the complete dataset. The thick curve shows the local mean of the modes, whereas the thin curves show the standard deviation for each sample size. Four speakers (cf. the curves with "additional" peaks in the rightmost panel of Figure 9) exhibited bimodal pitch distributions, and their primary modes do not seem to fully converge even after 3 minutes of speech is included. These speakers contribute to the secondary "row" of data points below the overall mode.

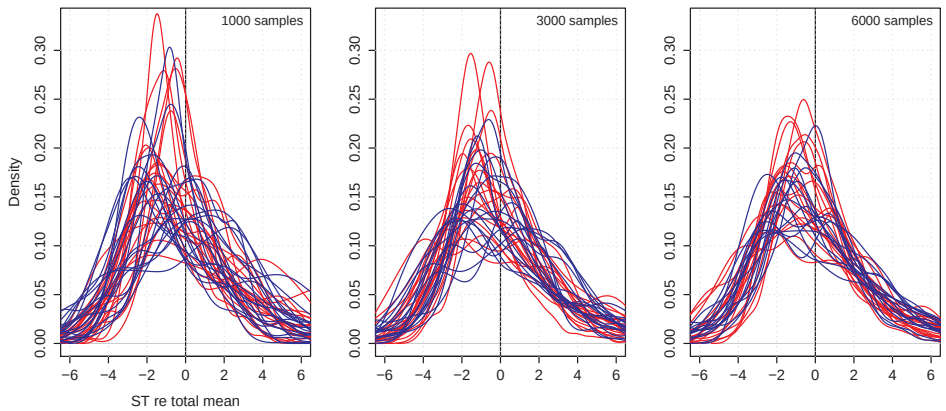


Figure 8: Distribution of a randomly selected subset of 1000, 3000 or 6000 consecutive pitch samples from 40 speakers. The pitch values are referred to the speaker-specific total **mean**, shown as the black vertical line in each plot.

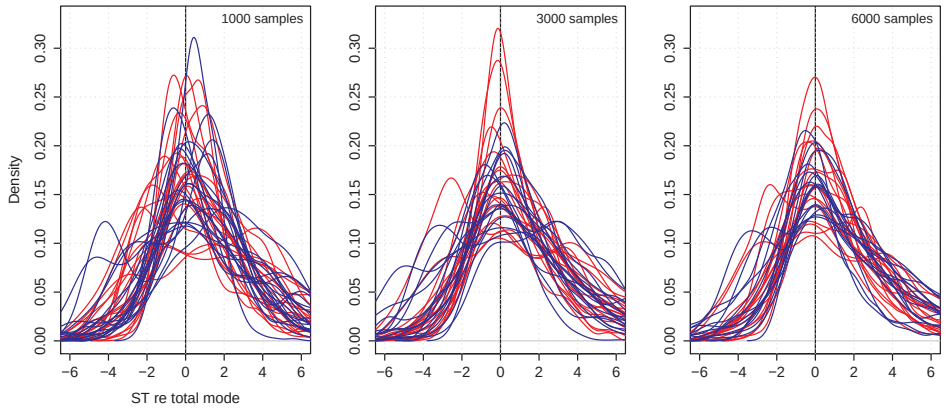


Figure 9: Distribution of a randomly selected subset of 1000, 3000 or 6000 consecutive pitch samples from 40 speakers. The pitch values are referred to the speaker-specific total **mode**, shown as the black vertical line in each plot. Male speakers are indicated with blue lines; females with red.

6 Conclusions

It was confirmed that in a sufficiently large dataset, a majority of the pitch values measured from each individual speaker tend to be distributed in a roughly similar fashion. This is likely to reflect the natural modes of vibration of the vocal folds and thus the pitch ranges of probable comfort vs. discomfort for the speaker. The primary distributions tend to be generally right-skewed. This observation is consistent with previous data (see, e.g., Traunmüller & Eriksson, 1995). The skewed distribution may be at least partly due to the fact that the length of the vocal folds sets a natural lower limit to glottal frequency, whereas humans can rather flexibly stretch their vocal folds in order to increase the pitch of their voices.

On the basis of these two corpora, it is typical for speakers to exhibit a primary pitch range that extends about 3–6 ST below and 6–12 ST above the pitch mode. Secondary "bulks" of data may be observed below and/or above the main range in the pitch distribution. In case these local modes occur at a distance of 12 ST (i.e., one octave) from the main pitch mode, it is to be suspected that they reflect a tendency of the speaker to use non-modal laryngeal settings (such as creaky voice or falsetto) and/or that the pitch analysis parameters have not been set in an optimal way for the speaker in question. For specific research purposes, it may be desirable to keep those results where the speaker's actual fundamental frequency has potentially been halved or doubled, since these may provide information about voice quality changes. In some cases, however, the additional modes may be due to other overlapping speakers or periodic background noise and need to be excluded. The present study paves the way for further research on the effects of various technical issues on pitch analysis, such as those of recording equipment, background noise, overlapping speech, voice quality differences, etc.

The minimum and maximum pitch do not provide a reliable summary of the speaker's preferred pitch range, since they are easily affected by non-modal voice

quality as well as by the selected analysis parameters. The standard deviation of the bootstrapped means and modes was reduced to less than 1 semitone after analyzing about 30 seconds or about 1500 pitch frames of net speaking time, given that the analysis parameters were set in an appropriate way. This may already be accurate enough for many research purposes. In case it is possible to determine the pitch mode of each particular speaker within a speech corpus, the mode is a good reference level for comparing the ways in which different speakers utilize their typical pitch ranges.

The tools for the analysis of pitch distributions may be applied in various domains, such as phonological models of intonation or clinical voice assessment. Given that some aspects in the pitch distributions may be highly speaker-dependent and relatively stable across different situations, the present tools may be applicable in the study of social identity (cf. Pierrehumbert et al., 2004; Munson, 2007; Cartei et al., 2014; Munson et al., 2015) and in the development of forensic speaker recognition (see Kinoshita et al., 2009). In terms of external factors that can affect speech, the tools for analyzing pitch distributions may be useful in studies of the effects of noise on speech production (cf. Hazan & Baker, 2011; Vainio et al., 2012) or for revealing whether speakers tend to accommodate their pitch levels to those of other speakers (cf., Gregory et al., 1993; 2001; Bosshardt et al., 1997; Babel and Bulatov, 2012; Garnier et al., 2013). Our findings will also be of interest in the analysis of the sequential unfolding of spoken social interaction, where the pitch range of the participants may systematically vary according to the position of a spoken turn within a larger sequence of turns (Stevanovic et al., submitted) and where speakers may be seeking to match each other's pitch levels according to sequential contingencies (Szczepek-Reed, 2010; Stevanovic & Lennes, submitted).

References

Babel M., & D. Bulatov D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55, 231–248.

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer* [Computer program]. Version 5.4.17, retrieved 3 September 2015 from http://www.praat.org/.

Bosshardt H.-G., Sappok, C., Knipschild, M., & Hölscher, C. (1997). Spontaneous imitation of fundamental frequency and speech rate by nonstutterers and stutterers. *Journal of Psycholinguistic Research* 26, 425–448.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, 25, 1044-1098. doi:10.1080/01690965.2010.504378.

Cartei, V., Cowles, W., Banerjee, R. & Reby, D. (2014). Control of Voice Gender in Pre-Pubertal Children. *British Journal of Developmental Psychology* 32(1): 100–106.

Couper-Kuhlen, E. (1996). The prosody of repetition. On quoting and mimicry. In: Elizabeth Couper-Kuhlen & Margret Selting (eds.), *Prosody in Conversation*. Cambridge University Press, Cambridge, 366-405.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23(2): 283–292.

Efron, B. (2003). Second Thoughts on the Bootstrap. *Statistical Science* 18(2): 135–140.

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

The FinINTAS Corpus of Spontaneous and Read-aloud Finnish Speech. URN http://urn.fi/urn:nbn:fi:lb-20140730194. [Speech corpus].

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., Emanuel A. Schegloff, S. A. Thompson (eds.), *Interaction and Grammar*. Cambridge: Cambridge University Press: 134-84

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural Correlates of Phonetic Convergence and Speech Imitation. *Frontiers in Psychology* 4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3769680/, accessed June 27, 2015.

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication,* 40, 189–212.

Gregory, S. W., Green, B. E., Carrothers, R. M., & Dagan, K. A. (2001). Verifying the primacy of voice fundamental frequency in social status accommodation. *Language and Communication,* 21, 37–60.

Gregory, S. W., Webster, S., & Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language and Communication*, 13, 195–217.

Hazan, V. & Baker, R. (2011) Acoustic-Phonetic Characteristics of Speech Produced with Communicative Intent to Counter Adverse Listening Conditions. *The Journal of the Acoustical Society of America,* 130(4), 2139–2152.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America* 117(4): 2193-200.

Kaimaki, M. (2010). Tunes in free variation and sequentially determined pitch alignment: evidence from interactional organization. *Journal of Greek Linguistics*, 10/2: 213-250.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech Language and the Law,* 16(1), 91-111.

Ladd, D. R. (1996). *Intonational phonology.* Cambridge Studies in Linguistics 79.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, 11, 373–382.

Lennes, M. (2007). On pitch and perceptual prominence in conversational Finnish speech. *Proceedings of the International Congress of Phonetic Sciences 2007, 6.-10.8.2007, Saarbrücken, Germany*, 1061-1064.

Lennes, M. (2016). pitch-distributions: Version 1.3. doi:10.5281/zenodo.45868

Lennes, M., Aalto, D., & Palo, P. (2008). Puheen perustaajuusjakaumat: Alustavia tuloksia. In: O'Dell, M. L., & Nieminen, T., eds., *Fonetiikan päivät 2008. Tampere Studies in Language, Translation and Culture, Series B,* 3, 147-155. Tampere: Tampere University Press.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America,* 102, 1864–1877.

Munson, B. (2007) The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation. *Language and Speech*, 50(1), 125–142.

Munson, B., Crocker, L., Pierrehumbert, J. B., Owen-Anderson, A., & Zucker K. J. (2015). Gender Typicality in Children's Speech: A Comparison of Boys with and without Gender Identity Disorder. *The Journal of the Acoustical Society of America*, 137(4), 1995–2003.

Ogden, Richard (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association,* 31(1), 139–152.

Persson, R. (2013). Intonation and sequential organization: Formulations in French talk-in-interaction. *Journal of Pragmatics,* 57, 19-38.

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The Influence of Sexual Orientation on Vowel Production (L). *The Journal of the Acoustical Society of America*, 116(4), 1905–1908.

R Core Team, 2014. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.from http://www.R-project.org/.

Scherer, Klaus R., London, H., Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7(1): 31-44.

Stevanovic, M., Himberg, T., Niinisalo, M., Peräkylä, A., Sams, M. & Hari, R. (submitted).

Sequential approach to interpersonal synchrony: The case of joint decision-making.
Stevanovic, M. & Lennes, M. (submitted). Pitch matching – absolute or relative? On prosodic orientation across speaker changes.
Szczepek-Reed, B. (2004) Turn-final intonation revisited. In E. Couper-Kuhlen, & C. Ford (eds.): *Sound patterns in interaction: Cross-linguistic studies from conversation*. Amsterdam: John Benjamins. 97-117.
Szczepek-Reed, B. (2010). Prosody and alignment: a sequential perspective. *Cult Stud of Science Education*, 5, 859-867.
Titze, I. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85, 1699–1707.
Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript. Retrieved 3 September 2015 from http://www2.ling.su.se/staff/hartmut/f0_mandf.pdf.
Vainio, M., Aalto, D., Suni, A., Arnhold, A., Raitio, T., Seijo, H., Järvikivi, J., & Alku, P. (2012). Effect of Noise Type and Level on Focus Related Fundamental Frequency Changes. In *Interspeech 2012 – 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA.
Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *TRASP'2013*, Aix-en-Provence, France. 7-10.
Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3), 315-337.

SPPAS - MULTI-LINGUAL APPROACHES TO THE AUTOMATIC ANNOTATION OF SPEECH

Brigitte Bigi
Laboratoire Parole et Langage, CNRS, Aix-Marseille Université
5 avenue Pasteur, 13100 Aix-en-Provence, France
e-mail: brigitte.bigi@lpl-aix.fr

Abstract

The first step of most acoustic analyses unavoidably involves the alignment of recorded speech sounds with their phonetic annotation. This step is very labor-intensive and cost-ineffective since it has to be performed manually by experienced phoneticians during many hours of work.

This paper describes the main features of SPPAS, a software tool designed for the needs of automatically producing annotations of speech at the level of utterance, word, syllable and phoneme based on the recorded speech sound and its orthographic transcription. In other words, it can automatize the phonetic transcription task for speech materials, as well as the alignment task of transcription with speech recordings for further acoustic analyses.

Special attention will be given to the methodology implemented in SPPAS, based on algorithms which are as language-and-task-independent as possible. This procedure allows for the addition of new languages quickly and for the adaptation of this tool to the user's specific needs. Consequently, the quality of the automatic annotations is largely influenced by external resources, and the users can modify the process as needed. In that sense, phoneticians need automatic tools and these tools can be significantly improved by phonetician input.

Keywords: automatic, annotation, speech segmentation, multilingual, methodology

1 Introduction

Corpus annotation “can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process” (Leech, 1997). Annotation of speech recordings is relevant for many sub-fields of linguistics such as phonetics, prosody, gesture analysis or discourse studies. Corpora are annotated with detailed information at various linguistic levels, often with the use of specialized annotation software. As *large multimodal* corpora become prevalent, new annotation and analysis requirements are emerging. In order to be useful for purposes such as qualitative or quantitative analyses, the annotations must be time-synchronized (time-aligned). Temporal information makes it possible to describe behavior or actions of different subjects that happen at the same time, and time-analysis of multi-level

annotations can reveal levels of linguistic structures. Generally, “different annotation tools are designed and used to annotate the audio and video contents of a corpus that can later be merged in query systems or databases” (Abuczki & Baiat Ghazaleh, 2013). A number of software programs for manual annotation and analysis of audio and/or video recordings are available such as Transcriber (Barras et al., 2001), Praat (Boersma & Weenink, 2001), or Elan (Wittenburg et al., 2006), to name but just some popular ones that are both open-source and multi-platform.

In the past, phonetic study was mostly based on limited data. Currently, phonetic models are often expected to be built based on the acoustic analysis of large quantities of speech data supported with valid statistical analyses. The first step of most acoustic analyses unavoidably involves the alignment of recorded speech sounds with its phonetic annotation. This step is very labor-intensive and cost-ineffective since it has to be performed manually by experienced phoneticians requiring many hours of work. For speech engineers, this labor-intensive task can be assisted by computer programs. A number of free toolkits are currently available which can be used to automate the task, including the HTK Toolkit (Young & Young, 1993), Sphinx (Lamere et al., 2003), or Julius (Lee et al., 2001). In recent years, the SPPAS software tool has been developed to automatically produce “annotations which include utterance, word, syllabic and phonemic segmentation from a recorded speech sound and its transcription” (Bigi, 2012). In other words, this software can automatize the phonetic transcription task for speech materials, as well as the alignment task of matching transcriptions to the speech recordings for further acoustic analyses. SPPAS includes resources for various languages such as English, French, Italian, Spanish, and Mandarin Chinese. An important feature is that SPPAS is specifically designed to be used directly by linguists (not necessarily skilled in programming) in conjunction with other tools for the analysis of speech. It is a *free software*, as defined by Richard Stallman (2002), and distributed under the terms of the GNU Public License.

Modern technology gives linguists the means of refuting theories and models with large quantities of language data. In order to efficiently use annotation software, particularly for automatic annotations, a rigorous methodology is necessary. Section 2 of this paper presents how to collect a large set of time-aligned annotations for various domains or levels: orthographic transcription (time-aligned at the level of inter-pausal units), phonetics (words, syllables, phonemes), prosody (Momel and INTSINT), morpho-syntax (categories, groups), discourse (repetitions) and gestures. Some are annotated manually and most of them are generated automatically. The main features of SPPAS are presented together with the basic guidelines for its integration within such a framework. Section 3 describes the automatic annotations implemented in SPPAS, with algorithms as language-and-task-independent as possible. This allows adding new languages with a significant reduction of time compared to the development of such tools from scratch, because adding a new language in SPPAS only consists of adding the resources related to the annotation (like lexicons, dictionaries, models, sets of rules, etc). Consequently, the quality of the automatic

annotations is largely influenced by such resources, and phoneticians can contribute to improve them.

2 Introducing SPPAS in a corpus construction methodology

This section illustrates the kind of process for development of a corpus that contains rich and broad-coverage of multimodal/multi-level annotations. This involves a rigorous framework to ensure compatibilities between accurate annotations and time-saving methodologies. Indeed, "when multiple annotations are integrated into a single data set, inter-relationships between the annotations can be explored both qualitatively (by using database queries that combine levels) and quantitatively (by running statistical analyses or machine learning algorithms)" (Chiarcos, 2008). The expected result is time-aligned data, for all annotated levels including phonetics, prosody, gestures, syntax, discourse (cf. Figure 1). The wide range of annotations is costly to collect and annotate, both in terms of time and money. Consequently, each annotation that *can* be done automatically *must* be done automatically, because revising is expected to be less time-consuming and easier than annotating, as shown for example by the use of SPPAS in Yu (2013). Fortunately, the current state-of-the-art in computational linguistics allows many annotation tasks to be semi- or fully- automated. Unfortunately, the lack of interoperability between automatic annotation tools/data and manual annotation tools/data is still a challenge. Thus, despite the advances that have been achieved for annotating and analyzing corpora, many annotation frameworks and/or models for the construction and analysis of multimodal data continue to rely on "low-tech" and/or manual technologies.

In recent years, many annotation software/tools have become available for annotation of audio-video data. For a researcher looking for an annotation software tool, it might be difficult to select the most appropriate one. The choice of the software determines the annotation framework and that will be utilized and this process should be done carefully and *before* the creation of the corpus. To decide about usefulness and usability of a software, it is advisable to consider the issues listed below.

- The software license: the preference is for free and open source software. Even if a user can personally afford to pay for a license, he/she may wish to share his/her methodology with other students or researchers who cannot afford to buy it.
- The ease of use: the first, preference is for multi-platform software. Different scientific communities tend to use MacOS, Windows or Unix platforms. Multi-platform software makes sharing between such communities much easier. Secondly, usable software is preferred. A need to request help from an engineer each time a user needs to use a piece of software may pose a serious limitation.
- The strengths/weaknesses for specific annotation purposes. Users should investigate if the software has been found to be reliable and is likely to

improve the efficiency of annotation workflow, by either accelerating the work or enabling one to deal with more extensive data, or both.

- The type of data or analysis the tool/software is specifically designed to complete.
- The software compatibility with other annotated data, i.e. the availability of files to be imported/exported from/to several other data formats.

Before using any automatic annotation tool/software, it is important to consider its error rate (where applicable) and to estimate how those errors will affect the purpose for the annotated corpora.

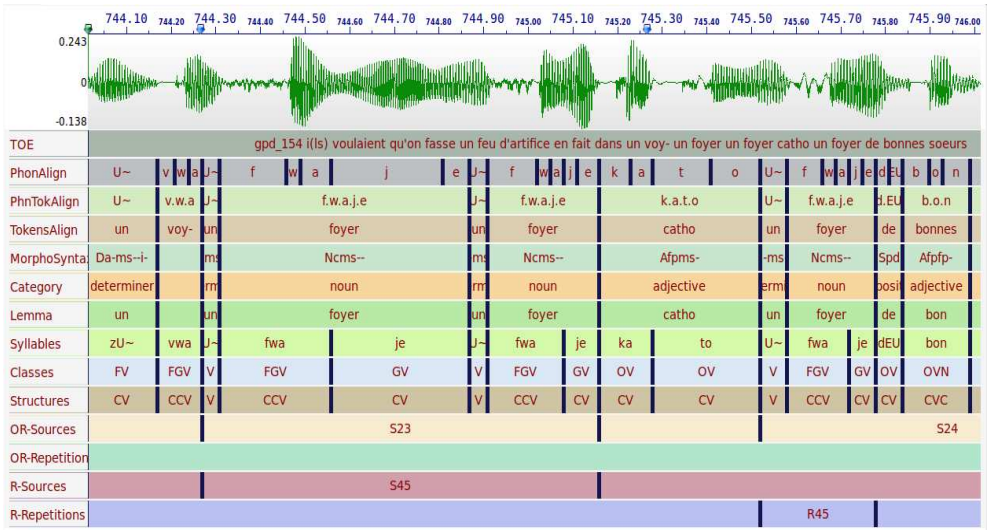


Figure 1: A selection of multi-level annotations based on the speech signal. The tier "TOE" is the enriched orthographic transcription and it was manually annotated. The other tiers were automatically annotated by SPPAS and MarsaTag (Rauzy, 2014) software.

In the following part of this section, we very briefly introduce selected annotation software programs that were included as part of the proposed annotation methodology: Praat, Elan and SPPAS.

Praat is a tool for manually annotating sound files. It provides different visualizations of audio data - waveform or spectrogram display - and, among other things, enables pitch contour as well as formant calculation and visualization. The annotation files are in several Praat-specific ASCII formats, but Praat doesn't support any import or export to other formats. Fortunately, Praat-TextGrid file format is well-known in the community and external converters exist.

Elan is a tool for the creation of complex annotations for video (and audio) resources. Annotations can be created on multiple layers that can be hierarchically interconnected and can correspond to different levels of linguistic analysis. It also includes an advanced

search system. The annotation files are in a specific XML format, and Elan can import from and export to a variety of other formats, including Praat-TextGrid.

SPPAS is an annotation software that allows one to *automatically* create, visualize and search annotations of audio data. In fact, the analysis of the phonetic entities of speech nearly always requires the alignment of the speech recording with a phonetic transcription of the speech. This task is extremely labor-intensive - it may require several hours even for an experienced phonetician to transcribe and align manually a single minute of speech. It is thus obvious that transcribing and aligning several hours of speech by hand is not generally something which can be accomplished with ease. Therefore, among others, SPPAS includes automatic segmentation of speech. It offers a fully-automatic or semi-automatic annotation process, with a procedure outcome report to help the user in understanding particular steps. Some special features are offered in SPPAS for managing corpora of annotated files; e.g., a component to filter multi-level annotations (Bigi & Saubesty, 2015). Some other components are dedicated to the analysis of time-aligned data; as for example to estimate descriptive statistics, a version of Time Group Analyzer (Gibbon 2013), etc. SPPAS annotation files are in a specific XML format, and annotations can be imported from and exported to a variety of other formats, including Praat (TextGrid, PitchTier, IntensityTier), Elan (eaf), Transcriber (trs), Annotation Pro (antx) (Klessa et al., 2013, Klessa, 2015), Phonedit (mrk) (Teston et al., 1999), Sclite (ctm, stm), HTK (lab, mlf), subtitles formats (srt, sub) and CSV files. SPPAS can be used either with a Command-line User Interface or a Graphical User Interface as shown in Figure 2. So, there's no specific difficulty when using this software. The only potential brake on its usage is the need to integrate it in a rigorous methodology for the corpus construction and annotations.

The kind of process for obtaining rich and broad-coverage of multimodal/multi-levels annotations of a corpus is illustrated in Figure 3. It describes each step of corpus creation and annotation workflow. This Figure must be read from top to bottom and from left to right, starting with the recordings and ending with the analysis of annotated files.

After recording speech samples, the first step to perform is **IPUs segmentation**. IPUs (Inter-Pausal Units) are blocks of speech bounded by silent pauses of more than *X* ms (the *X* duration depends on the language; for French, the duration of 200 ms is commonly used), and time-aligned on the speech signal. IPUs segmentation should be verified manually. The outcome of this automatic procedure depends on the quality of the recording: the better the quality, the better IPUs segmentation.

Orthographic transcription is often the minimum obligatory requirement for a speech corpus, as it is the entry point for most of the automatic annotations, including automatic speech segmentation. As a consequence, high quality orthographic transcription implies:

- high quality phonetic transcription,
- thus, high quality time-alignment of phonemes and tokens,
- thus, high quality syllabification,
- and so on.

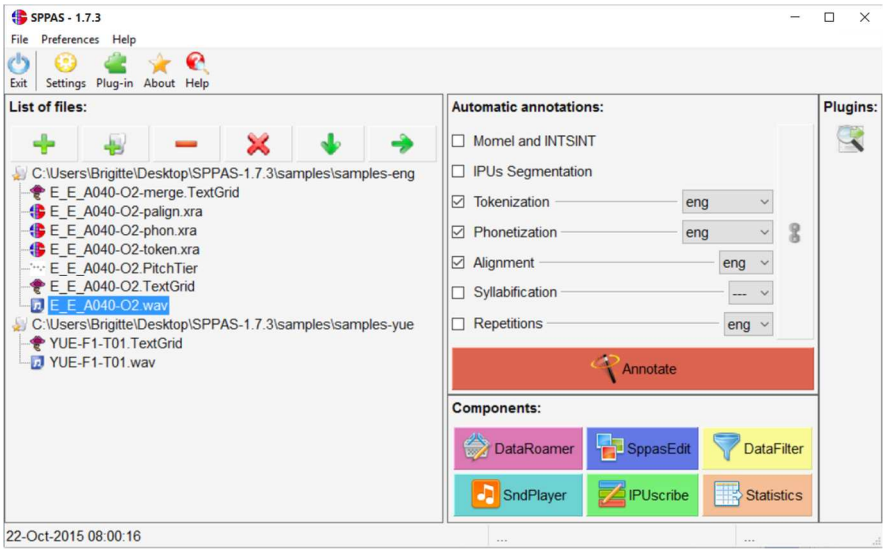


Figure 2: SPPAS Graphical User Interface. The left part indicates the list of files to work with; the middle part displays the functionalities of SPPAS (top: the whole list of automatic annotations; bottom: a set of 6 components provided to manage annotated data) and right part is dedicated to plug-ins (only one on this picture).

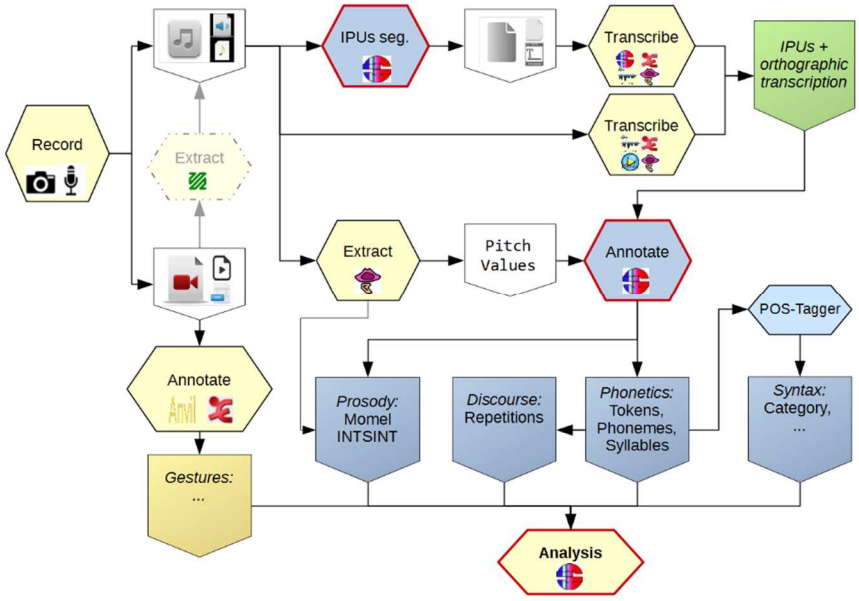


Figure 3: A multi-level corpus creation and annotation workflow. Yellow boxes represent manual annotations, blue boxes represent automatic ones.

The question then arises: what is “the better” orthographic transcription method? First, one of the characteristics of speech is the important gap between a word's

phonological form and its phonetic realizations. Specific realizations due to elision or reduction processes often occur and the same happens for other types of phenomena such as non-standard elisions, substitutions or addition of phonemes, noises, and laughter. Numerous studies have been carried out on prepared speech, such as broadcast news. However, conversational speech refers to a more informal activity, in which participants constantly need to manage and negotiate turn-taking, topic, etc. "on line" without any preparation which results in an even greater number and wider variety of non-standard events. Table 1 reports on the amount of such phenomena taken from three manually annotated samples of the following French corpora:

1. AixOx, read speech of short texts (Herment et al. 2012);
2. Grenelle II, a discourse at the French National Assembly (Bigi et al., 2012);
3. CID - Corpus of Conversational Data, spontaneous dialogs (Bertrand et al., 2008).

Table 1: Description of events in three different corpora available at http://sldr.org/sldr000786

	AixOx	**Grenelle II**	**CID**
Duration of the samples	0137s	0134s	0143s
Number of speakers	0004	0001	0012
Number of phonemes	1744	1781	**1876**
Number of tokens	1059	550	**1269**
Short silent pauses	0023	00**28**	0010
Filled pauses	0000	0005	00**21**
Noises (breathes, …)	000**8**	0000	0000
Laughter	0000	0000	000**4**
Truncated words	0002	0001	000**6**
Optional liaisons	0002	000**5**	0004
Elisions (non standard)	0021	0034	00**60**
Specific pronunciations	0037	0023	00**58**

These events may create obstacles for the automatic annotation process. Thus, SPPAS includes the support of an Enriched Orthographic Transcription (EOT). Here, transcribers are asked to indicate: filled pauses, short pauses, repeats, truncated words, noises, laughter, irregular elisions and specific pronunciations. These specific phenomena have a direct influence on the automatic phonetization procedure as shown in Bigi (2012).

The **Phonetics (Tokens, Phonemes, Syllables)** component of the workflow involves the process of taking the phonetic transcription text of an audio speech segment, like IPUs, and determining where particular phonemes occur in this speech segment. In SPPAS, this problem is clearly divided into three sub-tasks: Task 1 is tokenization, also called text normalization, Task 2 is phonetization, also called grapheme to phoneme conversion, and Task 3 is time-alignment, which is the speech segmentation task itself. All three sub-tasks are fully-automatic, but each annotation output can be manually

checked if desired (a semi-automatic mode). The current version of SPPAS (1.7.4) includes data and models for: French, English, Italian, Spanish, Catalan, Portuguese, Polish, Mandarin Chinese, Cantonese, Taiwanese and Japanese. The time-alignment of tokens (usually words) can be automatically derived from the time-alignment of phonemes. Afterwards, the time-alignment of *syllables* is derived from the time-alignment of phonemes using a rule-based system (Bigi et al., 2010).

In the **Discourse** domain, as shown in Figure 3, the time-alignment of tokens can also be used by SPPAS to automatically identify self-repetitions and other-repetitions (OR). This system is based only on lexical criteria to determine whether a token (only word in that case) is repeated or not. A set of rules are then fixed to filter such occurrences and to select only the relevant ones (Bigi et al., 2014). This system was used to propose a lexical characterization of OR: various statistics were estimated on the detected OR from CID corpus. It was also used to analyze if the same speech implies the same or different gestures in Tellier et al. (2012).

In the **Syntax** domain, a stochastic parser can be adapted to automatically generate morpho-syntactic and syntactic annotations. Actually, it must be adapted in order to account for the specifics of speech analysis, and to take time-aligned tokens as input. For French, MarsaTag (Rauzy, 2014) is available and can be used as a plugin of SPPAS.

The **Prosody** domain can also be investigated and included as part of the framework. Momel (Hirst & Espesser, 1993) is an example of a freely available algorithm for automatic modeling of fundamental frequency (*f0*) curves using a technique called asymmetric modal quadratic regression. This technique makes it possible to factor an *f0* curve into two components by an appropriate choice of parameters:

1. a macroprosodic component represented by a quadratic spline function defined by a sequence of target points <ms,Hz>.
2. a microprosodic component represented by the ratio of each point on the F0 curve to its corresponding point on the quadratic spline function.

INTSINT (an INternational Transcription System for INTonation) assumes that pitch patterns can be adequately described using a limited set of tonal symbols, T, M, B, H, S, L, U, D (standing for: Top, Mid, Bottom, Higher, Same, Lower, Up-stepped, Down-stepped respectively). Each one of these symbols characterizes a point on the fundamental frequency curve. Momel and INTSINT are tools enabling automatic annotations and are available as a Praat plug-in (Hirst, 2007), and re-implemented within SPPAS.

Gestures annotation can also play an important role in an annotation workflow, by reflecting the multimodal aspects of speech communication, however, this factor will not be described further in this paper. One can refer to Tellier (2014) for methodological insight into gesture annotation.

To sum up, this section presented a methodology for the annotation of recordings, based on both manual annotations and on annotations produced automatically with SPPAS, as illustrated in Figure 3. This methodology was established in the annotation of the CID - Corpus of Interactional Data (Bertrand et al., 2008; Blache et al., 2010),

and SPPAS was initially created to generate annotations only on the level of Phonetics. Subsequently, several other corpora were created using SPPAS in the context of various projects, e.g.: Amennpro (Herment et al., 2012), Cofee (Gorish, 2014), Multiphonia (Alazar et al., 2012), Typaloc (Bigi et al., 2015), and Variamu (Bigi & Fung; 2015). In order to meet new expectations and new project requirements, SPPAS was improved and extended with new functionalities and components. The proposed methodology has demonstrated flexibility as well as effectiveness and reliability in the demanding, real-world situations of corpora creation.

3 SPPAS: multi-lingual approaches

3.1 Text normalization

The first task faced by any Natural Language Processing system is the conversion of input text into a linguistic representation. Digital written texts contain a variety of "non-standard" entry types such as digit sequences, acronyms and letter sequences in all capitals, mixed case words, abbreviations, Roman numerals, URL's and e-mail addresses. Speech transcriptions also contain truncated words, orthographic reductions, etc. Normalizing or rewriting such texts using ordinary words is an important issue for various applications. There is a greater need for work on text normalization, as it forms an important component of all areas of language and speech technology. Text normalization development is commonly carried out specifically for each language and/or task even if this work is laborious and time consuming. Actually, for many languages there has not been any concerted effort directed towards text normalization. Considering the above, as well as the context of genericity, producing reusable components for language-and-task-specific development is an important goal. This section describes SPPAS text normalization and concentrates on the aspects of methodology and linguistic engineering which serve to develop this multi-purpose multi-lingual text corpus normalization method.

SPPAS implements a generic approach, i.e. a text normalization method as *language and task independent* as possible. This enables adding new languages quickly when compared to the development of such tools from scratch. This method is implemented as a set of modules that are applied sequentially to the text corpora. The portability to a new language consists of inheriting all language independent modules and rapid adaptation of other language dependent modules. In the same way, for a new task, a module can be inherited from general processing modules, and adapted rapidly to create other specific modules.

The first step is to determine which modules to use, some are shared (the modules which do not depend on the language), and some are variable modules (language-dependent modules). This splitting and specification of work is really important. For modeling a new language, the shared modules will be inherited and the variable modules will be adapted to that language. It will economize the time needed to complete corpus normalization. The key idea is to concentrate the language knowledge in a set of lexicons and to develop modules which implement rules to deal with the knowledge elements. Shared modules are listed below:

- *Basic unit splitting module:* a segmentation module based on white spaces for Romanized languages and character-based for the other languages.
- *Replacing module:* implements a dictionary look-up algorithm to replace a string by another one. It is mainly used to replace special symbols like ° (degrees), for example.
- *Lowerize module:* used to convert the character-case.
- *Word-tokenization module:* fixes a set of rules to segment strings including punctuation marks for Romanized languages. This algorithm splits strings into words on the basis of a dictionary and a set of manually established rules. For example, in French “trompe-l'oeil” (*sham*) is an entry in the vocabulary and it will not be segmented. On the other hand, an entry like “l'oeil” (*the eye*) occurring in another context will be segmented into two separate words.
- *Sticking module* implements an algorithm to concatenate strings (or characters) into words based on a dictionary with an optimization criteria: *longest matching*.
- *Removing module* can be applied to remove strings of a text. The list of strings to remove is defined in a separate file. For certain applications, it is relevant for example to remove punctuation marks.

Apart from the abovementioned shared modules, SPPAS also includes several language-specific modules. One of them is the optional *number to letter module*. For example, the number “123” is normalized as “one_hundred_twenty-three” for English and “ciento_veintitres” in Spanish. It is thus necessary to implement this module for each new language if numbers are used in the orthographic transcription. Adding a new language only consists of adding the list of tokens in the appropriate directory of the SPPAS package, and eventually writing the number to letter conversion. It means also that any phonetician can edit/modify the lexicon to get the expected result.

Another specific module has been developed to deal with enriched orthographic transcriptions. From the manual EOT (Enriched Orthographic Transcription), two types of transcriptions are automatically derived by the tokenizer: the “standard transcription” (a list of orthographic tokens/words) and the “faked transcription” that is a specific transcription from which the obtained phonetic tokens are used by the phonetization system. The following example illustrates an utterance text normalization extracted from the CID corpus in French:

Transcription: j'ai on a j'ai p- (en)fin j'ai trouvé l(e) meilleur moyen c'était d(e) [loger,locher] chez des amis (*I've we've I've - well I found the best way was to live in friends' apartment'*)
Standard transcription: j' ai on a j' ai p- enfin j' ai trouvé le meilleur moyen c'était de loger chez des amis
Faked transcription: j' ai on a j' ai p- fin j' ai trouvé l meilleur moyen c'était d locher chez des amis

The standard one is “human-readable” and can be used for further processing by any automatic system, e.g., an automatic syntax analysis. The faked one is useful

mainly for the grapheme-to-phoneme conversion system. In the case of standard orthographic transcription instead of EOT, both the generated standard and faked transcriptions are identical. See Bigi et al. (2012) for an evaluation of the impact of such EOT on the automatic phonetization system of SPPAS.

We applied the SPPAS automatic tokenizer on the 16 files of the French CID corpus, which were fully transcribed with EOT. Each file represented the transcription of one hour of speech in the context of eight dialogues. This process was accomplished in 95s with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the normalized text (a total of 120,000 tokens) including standard and faked transcriptions.

3.2 Phonetization

Phonetic transcription of text is an indispensable component of text-to-speech systems and is used in acoustic modeling for automatic speech recognition and other natural language processing applications. Generally, grapheme-to-phoneme conversion is a complex task, for which a number of diverse solutions have been proposed. It is a structure prediction task; since both the input and output are structured, consisting of sequences of letters and phonemes, respectively. It can be implemented in many ways, often roughly classified into dictionary-based and rule-based strategies, although many intermediate solutions exist. In the context of our study, the phonetization process takes the normalized transcription of the speech signal as input and produces the supposed pronunciation. The phonetization of speech corpora requires a sequence of processing steps and resources in order to convert the normalized text into its constituent phones.

SPPAS implements a dictionary-based approach, which is relatively language-independent. The dictionary includes phonetic variants that are proposed for the aligner to choose the phoneme string. The hypothesis is that the answer to the phonetization question can be found in the speech signal. Consequently, an important step is to build the pronunciation dictionary, where each word in the vocabulary is expanded into its constituent phones, including pronunciation variants. Depending on the language, the availability of such resources varies. In the SPPAS data set, the dictionary includes a large set of entries for English, French, Italian, Polish, an acceptable number of entries for Catalan, Mandarin Chinese, Spanish, Japanese, Cantonese, and a rather poor number of entries for Taiwan Southern Min. In addition, SPPAS implements an algorithm for phonetization of unknown words (e.g., proper names, speech reductions or mispronunciations). The present grapheme-to-phoneme conversion system is based on the idea that given enough examples it should be possible to predict the pronunciation of unseen words purely by analogy. The system is then applied to missing words during the phonetization process (and not during a training stage), and is only based on knowledge provided by the dictionary. The algorithm consists of exploring the unknown entry from left to right, then right to left, to find the longest strings in the dictionary. Since SPPAS-Phonetization only uses the pronunciation dictionary either for known or unknown words, the quality of such an annotation depends mainly on the quality of a particular resource. Another

consequence of such a system is that adding a new language in SPPAS-Phonetization only consists in adding the pronunciation dictionary in the appropriate directory of the SPPAS package. It also means that any phonetician can use their own dictionary.

We applied the SPPAS automatic phonetizer on the 16 normalized files of the French CID corpus. The process was accomplished in 71s with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the phonetized transcription, including pronunciation variants.

3.3 Speech segmentation

Phoneme alignment is the task of proper positioning of a sequence of phonemes in relation to a corresponding continuous speech signal. In the alignment task, we are given a speech utterance along with the given phonetic representation for that utterance. Our goal is to generate an alignment between the speech signal and the phonetic representation. Manual alignment has been reported to take between 11 and 30 seconds per phoneme (Leung and Zue, 1984). An automatic time-alignment system is then essential for the annotation of large corpora.

SPPAS is based on the use of the Julius Speech Recognition Engine (Lee et al., 2001). This choice is motivated by four main reasons:

1. the Julius toolkit is open-source, so there is no specific reason to develop a new one;
2. it is easy to install which is important for end-users;
3. it's usage is relatively easy so it was convenient to integrate it in SPPAS;
4. its performance corresponds to the state-of-the-art of other available systems of such kind.

The Julius alignment task processes in two-steps: The first step selects the phonetization and the second step performs the segmentation. A finite state grammar that describes sentence patterns to be recognized and an acoustic model are needed. This grammar essentially defines constraints on what the Speech Recognition Engine can expect as input. SPPAS generates the grammar automatically from the phonetized files. Speech alignment also requires an acoustic model in order to align speech. This involves a file that contains statistical representations of each of the distinct sounds in a language. The original Julius distribution only includes Japanese acoustic models. However since it can use acoustic models of HTK-ASCII format (a common format used by many systems), this system can also be adapted to other languages. Consequently, any user can train it's own acoustic model, or get it from the web, and integrate it in SPPAS.

Most of the acoustic models already included in SPPAS were trained by the author of this paper with HTK by taking a training corpus of speech, previously segmented into utterances and phonetized. Ideally, the phones would have unique articulatory and acoustic correlates. But acoustic properties of a given phone can depend on the phonetic environment. These co-articulation phenomena motivated the adoption of context-dependent models such as triphones, for each language we had enough data for training. To train such acoustic models, the training procedure is based on the

VoxForge tutorial[1], except that VoxForge suggests using only word transcription as input, and we allow (and prefer) to use phonetized ones. The outcome of this training procedure is dependent on the availability of accurately annotated data and on good initialization. Acoustic models were trained from 16 bits, 16000 Hz wav files. This procedure had three main steps:

- data preparation,
- monophones generation,
- triphones generation.

Step 1 establishes the list of phonemes, plus silence and short pauses. It converts the input data (phonetization of the corpus) into an HTK-specific data format. It codes the (audio) data in a process known as "parameterizing the raw speech waveforms into sequences of feature vectors". Step 2 involves monophones generation. It creates a Flat Start Monophones model by defining a prototype model and copying this model for each phoneme. Then, this flat model is re-estimated using the provided data files to create a new model. Step 3 creates tied-state triphones. From our previous studies on French and Italian, we observed that five minutes of manually-time-aligned data are sufficient to train the initial model; and we found that about 10-30 minutes of manually-phonetized data are required to train a good monophone model. The orthographic transcription of several hours of speech will allow one to train a triphone model. As a consequence, any phonetician who had already created such a corpus for any language could share it privately with the author of SPPAS for a new acoustic model to be trained and publicly shared with the community.

We applied the SPPAS automatic aligner on the 16 audio files of the CID corpus, which were already converted to wav/mono/16000Hz/16bits, as the default in SPPAS. The process of time-aligning these 14000 IPUs was accomplished in 84min with SPPAS version 1.7.2 on a 2009-Desktop PC. The result was a set of 16 files containing the time-aligned phonemes and tokens (as shown in tiers 2, 3 and 4 of Figure 1).

3.4 Syllabification

The syllabification implemented in SPPAS is a rule-based system based on time-aligned phonemes. This phoneme-to-syllable segmentation system is based on two main principles:

1. a syllable contains a vowel, and only one;
2. a pause is a syllable boundary.

These two principles focus on the problem of finding a syllabic boundary between two vowels. Phonemes were grouped into classes and rules established to deal with these classes. We defined general rules as well as a small number of exceptions. Consequently, the identification of relevant classes is important for such a system. The rules follow usual phonological statements for most of the corpora and Romance

[1] http://www.voxforge.org

languages. An external configuration file indicates phonemes, classes and rules. This file can be edited and modified by any user to adapt the syllabification to any language or phoneme encoding. In the current version of SPPAS the respective sets of rules are available for French, Italian and Polish.

3.5 Self- and Other-repetitions

Other-repetition (OR) is a device involving the reproduction by a speaker of what another speaker has just said. Other-repetition has been found as a particularly useful mechanism in face-to-face conversation due to the presence of discursive or communicative functions. Among their various functions in discourse, repetitions serve the purpose of facilitating comprehension by providing less complicated discourse, while also establishing connection between various stages of discourse (cohesion), and also function as a device for getting or keeping the floor. SPPAS implements a semi-automatic method to retrieve other-repetition occurrences (Bigi et al., 2014). A key-point is that the proposed automatic detection is based on observable cues which can be useful for OR's identification from the time-aligned tokens. SPPAS captures repetitions which can be an exact repetition (named strict echo) or a repetition with variation (named non-strict echo). The rules of this system have been adapted to the detection of self-repetitions in the context of a study presented in (Tellier et al., 2012). As such, this method is intrinsically language-independent.

4 Conclusion

This paper described the automatic annotation systems included in SPPAS, a computer software tool designed and developed by the author to handle multiple language corpora and/or tasks with the same algorithms in the same software environment. Only the resources (e.g., dictionaries, lexicons, acoustic models) are language-specific and the approach is based on the simplest resources possible. The present work emphasizes new practices in the methodology of tool developments: considering the problems with a generic multi-lingual aspect, sharing resources, and putting the end-users in control of their own computing.

We hope this work will be helpful for the linguistic research community, and especially for those involved in speech research, as far as possible. Phoneticians are of crucial importance for resource development as they can contribute to improve the resources used by automatic systems. In the case of SPPAS, the improved software versions are systematically released to the public and serve to benefit of the whole community. Resources are distributed under the terms of a public license, so that SPPAS users have free access to the application source code and the resources of the software they use, free to share the software and resources with other people, free to modify the software and resources, and free to publish their modified versions of the software and resources.

References

Abuczki, Á., & Baiat Ghazaleh, E. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 1(9), 86-98.

Alazard, C., Astésano, C. and Billières, M. (2012). MULTIPHONIA: a MULTImodal database

of PHONetics teaching methods in classroom InterActions. *Language Resources and Evaluation Conference*, Istanbul (Turkey).
Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1), 5-22.
Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3), 1-30.
Blache, P., Bertrand, R. Bigi, B., Bruno, E., Cela, E., Espesser, R., Ferré, G., Guardiola, M., Hirst, D., Magro, E.-P., Martin, J.-C., Meunier, C., Morel, M.-A., Murisasco, E., Nesterenko, I., Nocera, P., Pallaud, B., Prévot, L., Priego-Valverde, B., Seinturier, J., Tan, N., Tellier, M., & Rauzy S. (2010). Multimodal Annotation of Conversational Data. The *Fourth Linguistic Annotation Workshop, ACL* 2010, 186-191, Uppsala, Suède.
Bigi, B. (2012). SPPAS: a tool for the phonetic segmentations of Speech. *The Eight international conference on Language Resources and Evaluation*, Istanbul (Turkey), 1748-1755,
Bigi, B., P. Péri, R. Bertrand (2012). Orthographic Transcription: Which Enrichment is required for Phonetization?, Language Resources and Evaluation Conference, Istanbul (Turkey), pages 1756-1763
Bigi, B. (2013). A phonetization approach for the forced-alignment task. 3rd Less-Resourced Languages workshop, 6th Language & Technology Conference, Poznan (Poland).
Bigi, B. (2014). A Multilingual Text Normalization Approach. *Human Language Technologies Challenges for Computer Science and Linguistics.* LNAI 8387, Springer, Heidelberg, 515-526.
Bigi, B., Bertrand R., & Guardiola, M. (2014). Automatic detection of other-repetition occurrences: application to French conversational speech. *9th International conference on Language Resources and Evaluation (LREC)*, Reykjavik (Iceland), 2648-2652.
Bigi, B., Watanabe, T., & Prévot, L. (2014). Representing Multimodal Linguistics Annotated Data. *9th International conference on Language Resources and Evaluation (LREC)*, Reykjavik (Iceland), 3386-3392.
Bigi, B., Klessa, K., Georgeton, L., & Meunier, C. (2015). A syllable-based analysis of speech temporal organization: a comparison between speaking styles in dysarthric and healthy populations. *INTERSPEECH*, Dresden (Germany), 2977-2981.
Bigi, B., & Saubesty, J. (2015). Searching and retrieving multi-levels annotated data. *Gesture and Speech in Interaction.* Nantes (France), 31-36.
Bigi, B., & Fung, R. (2015). Automatic Word Segmentation for Spoken Cantonese. *Oriental Chapter of International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques.* Shanghai (China).
Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10), 341-345.
Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., & M. Stede (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, 49(2), 271-293.
Gibbon, D. (2013). TGA: a web tool for Time Group Analysis, in Hirst, D.J., & Bigi, B. (Eds.): *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, 66-69.
Gorisch, J., Astésano, C., Gurman Bard, E., Bigi, B., & Prévot, L. (2014). Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. *9th International conference on Language Resources and Evaluation*, Reykjavik (Iceland), 2648-2652.
Herment, S., Tortel, A., Bigi, B., Hirst, D., & Loukina A. (2014). AixOx, a multi-layered learner's corpus: automatic annotation. Specialisation and Variation in Language Corpora. In Díaz-Negrillo, A., & Díaz-Pérez, F.J. (Eds*.): Linguistic Insights: Studies in Language and Communication.* 41-76.
Hirst, D.J., & Espesser, R. (1993). Automatic Modelling Of Fundamental Frequency Using A Quadratic Spline Function. *Travaux de l'Institut de Phonétique d'Aix*, 85, 75-85.

Hirst, D.J. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modeling and coding intonation. In *Proceedings of the XVIth International Conference of Phonetic Sciences*, Saarbrücken, August 2007, 1233-1236.

Klessa, K., Karpiński, M., & Wagner, A. (2013). Annotation Pro-a new software tool for annotation of linguistic and paralinguistic features. In *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, 51-54.

Klessa, K. (2015). Annotation Pro [Software tool]. Version 2.2.6.0. Retrieved from: http://annotationpro.org/ on 2015-05-19.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., & Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, vol. 1, 2-5.

Lee, A., Kawahara, T., & Shikano, K. (2001). Julius --- an open source real-time large vocabulary recognition engine. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, 1691-1694.

Leech, G. (1997). Introducing corpus annotation. In Garside, R., Leech, G., & McEnery, A.M. (Eds.): Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, 1-18.

Leung, H.C., & Zue, V.W. (1984). A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, vol. 9, 73-76.

Rauzy, S., De Montcheuil, G., & Blache, P. (2014). MarsaTag, a tagger for French written texts and speech transcriptions. *Second Asia Pacific Corpus Linguistics Conference*, Hong Kong.

Stallman R. (2002). Free Software, Free Society: Selected Essays of Richard M. Stallman. Retrieved on 2015-09-27 from: https://www.gnu.org/philosophy/fsfs/rms-essays.pdf

Tellier, M. (2014). Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés. Discours. *Revue de linguistique, psycholinguistique et informatique*, 15. https://discours.revues.org/8917

Tellier, M., Stam, G., & Bigi, B. (2012). Same speech, different gestures?, In *5th International Society for Gesture Studies (ISGS)*, Lund, Sweden.

Teston, B., Ghio, A., & Galindo, B. (1999). A multisensor data acquisition and processing system for speech production investigation. In *International Congress of Phonetic Sciences (ICPhS)*, 2251-2254.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC.* 1556-1559.

Young, S. J., and Sj Young (1993). The HTK hidden Markov model toolkit: Design and philosophy. University of Cambridge, Department of Engineering.

Yu, J. (2013). Timing analysis with the help of SPPAS and TGA tools. In Bigi, B., & Hirst, D. (Eds.): *Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France. 70-73.

USING WEB AUDIO TO DELIVER INTERACTIVE SPEECH TOOLS IN THE BROWSER

Mark Huckvale
Speech, Hearing and Phonetic Sciences
University College London
e-mail: m.huckvale@ucl.ac.uk

Abstract

In 2014, the number of web pages delivered to tablets and smartphones overtook the number delivered to laptop and desktop computers, with a majority of users saying they prefer these new portable platforms over conventional computers for many tasks. This shift in device use provides both opportunities and challenges for providers of speech analysis tools, phonetic demonstrations and language teaching aids. It is an opportunity because web standards mean we can make our applications available to a wide audience through a single consistent programming architecture rather than writing for one particular computing platform. It is a challenge because tablets and smartphones are less powerful, require different programming skills and have different limitations in terms of user interface.

In this article, I will show how interactive applications in Phonetics and Speech Science can be written to run in web browsers on any computing platform. These are native web applications, written in HTML, CSS and JavaScript that can capture, replay, display, process, and analyze audio using the Web Audio API without needing any plug-ins. I will describe - and give the URLs of - some demonstration applications. I will discuss some future opportunities in the area of collaborative research and some remaining challenges that arise from incompatibilities across browsers. My audience is teachers and students with intermediate web programming skills wanting to build custom speech displays, perform custom speech analysis or run speech audio experiments over the web.

Keywords: Speech audio, speech analysis, internet, web, programming

1 Introduction

There have been many changes in the field of computing since I started writing speech analysis software in the 1980s. The first Speech Filing System (SFS) tools (Huckvale et al, 1987) were written for the Unix operating system running on engineering workstations only available in scientific laboratories. But as personal computing grew, I developed and ported them to mainstream computing platforms: first to MS-DOS and then to Windows 3, 95, NT, XP, Vista, 7, 8 and now 10. By targeting one platform, my goal was to make the tools available to the largest number of people for the lowest cost in support. Other authors of speech tools have targeted Windows, Mac

OS, Linux or the Java VM, but all have primarily addressed users of desktop and laptop computers which were the descendants of those engineering workstations.

Recently however, the landscape of personal computing has changed radically. In 2014, it is said, more web pages were delivered to tablets and smartphones than were delivered to laptop and desktop computers. When asked, users say they prefer these new portable devices over conventional computing devices for a number of activities, including accessing the web[1], managing communications and consuming entertainment media. That preference is probably to do with portability, permanent network connectivity, and significantly better ease-of-use compared to laptops and desktops. In this landscape, our speech analysis tools look out of place, not only in terms of their restriction to particular desktop computing platforms, but because of their old-fashioned user interface and their need for installation and configuration.

There are gains to be had if we were able to make our tools compatible with modern tablets and smartphones by converting them to web applications. Our tools would become more widely available to a broader range of users; distribution would be simplified with our applications sitting on web pages and no longer needing installation, and by exploiting web standards, we would be programming for a single environment compatible with all computing platforms.

There are challenges too, of course. Our tools will need a user interface that doesn't require a mouse or keyboard which may involve re-thinking how they are operated – but the result may be tools which are more intuitive and easier to use by non-technical people. The available computational power and storage in tablets is less than in desktops (although improving every year) – but this can be addressed through the use of cloud computing, which also allows for more collaborative work. The personalization of tools with scripts might be more difficult for users – but we have the opportunity for an open plug-in architecture for analysis algorithms too.

In this paper, I look towards one approach to putting our speech analysis tools into the hands of modern users of tablets and smartphones: that of exploiting the industry standard programming development environment for audio processing available within web browsers. Web browser applications are different to smartphone and tablet "apps" in that typically they do not need installation or special privileges to operate and they can be delivered in the same way as ordinary web pages. Web applications are good for the novice developer in that the only tools needed to write them are a text editor and a browser. Also because all the program sources are available by default, this environment is more open to the sharing of code and algorithms. My goal is to provide practical information on how to build speech audio applications for the teacher or student wanting to build custom speech displays, perform custom speech analysis or run speech audio experiments over the web. My audience is intermediate

[1] http://www.statista.com/statistics/326100/most-important-device-for-connecting-to-the-internet-uk/

level developers who have already come to terms with basic elements of web programming.

2 The web software development environment

The browser application environment has special characteristics which provide a number of challenges for software development. The first is the separation between client and server: the client being the browser application running on the user's computer, and the server being the remote system that delivers web services, see Figure 1. Applications can be programmed to run solely on the client, solely on the server or on a mixture of the two. Typically security constraints limit what services an application can call on either server or client. Notably, the application has very limited access to data stored on the client or to the local hardware. This is to prevent remote applications taking control of the client's computer, such as recording audio or accessing personal information without permission. Additionally the communication between client and server can be unreliable – particularly in mobile networks – so applications need to be robust to slow network transfer speeds. In practice, this means that communications between client and server must be performed in the background, with applications still functional while data is being transferred, and they have to be written with this asynchrony in mind.

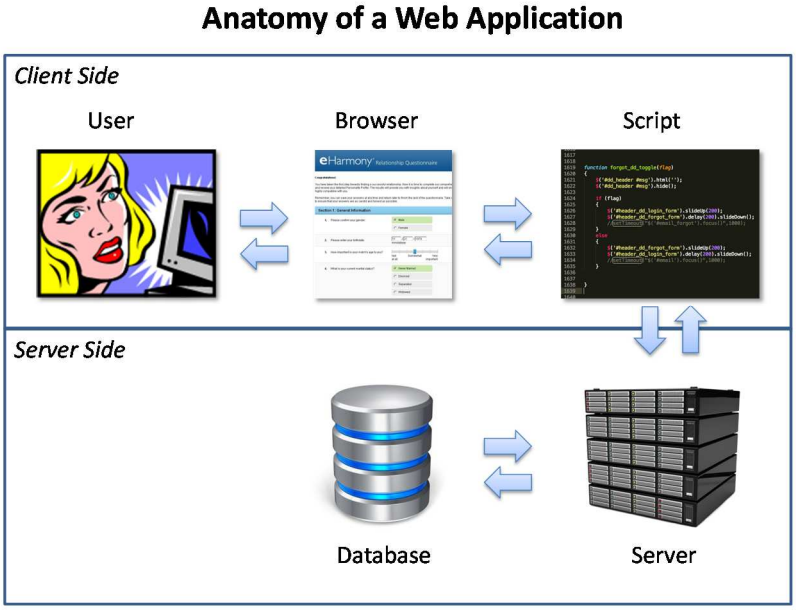


Figure 1: Anatomy of a web application

On the client side, the dominant programming framework involves HTML5, CSS and JavaScript. HTML5 is the mature content mark-up language for web pages, which gives structure to the information displayed in the browser. CSS is the styling

language which controls the layout and typography for that information, as well as controlling other graphical elements aspects of the page. JavaScript is a programming language which is able to manipulate elements of the web page, as well as performing general purpose programming tasks on the client, communicate with the server, and facilitate access to many other services provided by the browser (such as audio). The combination of HTML, CSS & JavaScript is also becoming the framework of choice for the development of smartphone and tablet “apps”, so knowledge of these is now even more important for the modern programmer.

On the server side, scripts may be written in a wide variety of languages, including C++, Python, Perl, and PHP as well as JavaScript. Typically server-side scripts are used to mediate access to databases – providing permanent data storage for transient client-side applications. In contrast to client-side scripts which are distributed in source form, server side scripts are not generally available to users, and this difference can be used to enforce security and ownership of intellectual property.

In the following sections, I will focus on the novel aspects of writing web applications that manipulate audio using the web audio API (application programming interface)[2]. In section 3, I give a complete simple demonstration of a web audio application, while in section 4, I introduce some more advanced capabilities.

3 Web audio demonstration

In this section, I give the source listing of a complete web audio application. This application loads an audio file from the client’s computer, displays the signal as a waveform and allows the user to replay the audio. It exploits the Flotr graphing library, which is described in section 5. Figure 2 shows the application running.

```
<html>
<head>
<meta charset="utf-8">
<title>WebAudio Demonstration</title>

<!—flotr graphics library from
http://www.humblesoftware.com/flotr2/ -->
<script type="text/javascript" src="flotr2.min.js"></script>

<script>

// audio context
"var context"=null;

// storage for signal
var signal=[];

// create audio context
function createContext()
{
   if (context==null) {
         // create the audio context
         try {
              context = new window.AudioContext();
```

[2] https://dvcs.w3.org/hg/audio/raw-file/tip/webaudio/specification.html

```
        }
        catch {
             alert('Web Audio API is not supported in this
browser.');
        }
   }
}

// display the waveform
function displayAudio()
{
   // get target container on page
   var container = document.getElementById("waveform");

   // squeeze signal into 2000 points for efficient plotting
   var  factor=Math.ceil(signal.length/2000);

   // load signal into graph
   var data = [];
   for (var i=0;i<signal.length;i+=factor) {
        var min=signal[i];
        var max=signal[i];
        for (var j=1;j<factor;j++) {
             if (signal[i+j] < min) min=signal[i+j];
             if (signal[i+j] > max) max=signal[i+j];
        }
        data.push([ i/context.sampleRate, min ]);
        data.push([ i/context.sampleRate, max ]);
   }

   // Draw Graph using Flotr library
   graph = Flotr.draw(container, [ data ], {
        title : "Waveform",
        shadowSize : 0,
        xaxis : {
             title : "Time (s)"
        },
        yaxis : {
             title : "Amplitude",
             titleAngle : 90
        },
        HtmlText : false
   } );

}

// load a file from client
function loadAudio()
{
   // get the filename
   var file = document.getElementById('filechoice').files[0];
   var filename = file.name;

   createContext();

   // set up a file reader
   var reader = new FileReader();

   reader.onload = functionI {
        var filedata = e.target.result;
        context.decodeAudioData(
             filedata,
             function onSuccess(buffer) {
                  // OK, take a copy of the samples
                  signal = new Array(buffer.length);
```

```
                    var srcbuf = buffer.getChannelData(0);
                    for (i=0;i<buffer.length;i++) signal[i] =
srcbuf[i];
                    // display waveform
                    displayAudio();
               },
               function onFailure() {
                    // load did not succeed
                    alert("decodeAudioData failed on "+filename);
               }
          );
   };

   reader.readAsArrayBuffer(file);
}

// play some audio
function playAudio()
{
   createContext();

   // create audio buffer source node
   sendsrc = context.createBufferSource();
   sendbuf =
context.createBuffer(1,signal.length,context.sampleRate);

   // copy in the signal
   senddat = sendbuf.getChannelData(0);
   for (i=0;i<signal.length;i++) senddat[i] = signal[i];

   // kick off replay
   sendsrc.buffer = sendbuf;
   sendsrc.loop = false;
   sendsrc.connect(context.destination);
   sendsrc.start(context.currentTime);
}

</script>
</head>
<body>

<h1>WebAudio Demonstration</h1>

<div style="height:1cm;width:100%;background-
color:lightgray;display:flex;
align-items:center;justify-content:center;margin-bottom:5mm;">

<input type="file" id="filechoice">
<button onclick="loadAudio()">Load Audio</button>
<button onclick="playAudio()">Play Audio</button>
</div>

<div style="height:10cm;width:100%;" id="waveform">
</div>

</body>
</html>
```

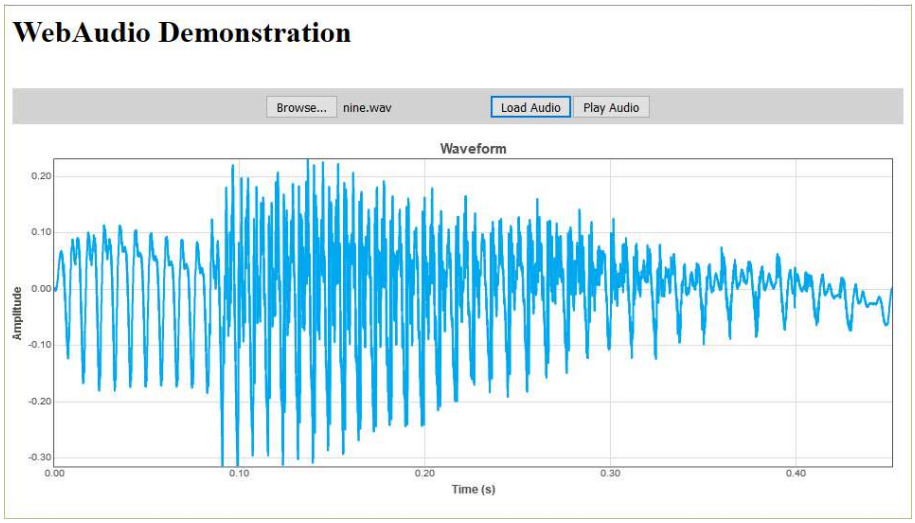


Figure 2. Web audio demonstration

Here is a brief commentary on some of the important elements of the demonstration.

At the heart of the audio functionality in modern web browsers is the AudioContext object. To access any of the web audio methods, it is necessary to first create an audio context object using the window.AudioContext method, as this code demonstrates:

```
try {
  context = new window.AudioContext();
  alert("context.sampleRate="+context.sampleRate);
}
catch {
  alert('Web Audio API is not supported in this browser.');
}
```

The sampleRate property of the AudioContext object gives the sampling rate for all audio operations in the browser. This is typically 44100 or 48000 samples per second. This cannot be changed, and the script must be written to work with the supplied rate.

To load an audio file from the client, a file input element needs to be placed on the web page for the user to select a particular file. For security reasons, scripts are not able to load files by pathname from the client machine. The file input HTML might look like this:

```
<input type="file" id="filechoice">
```

We can get access to the chosen file though the input element's files property.

To read the client file into the application, we can use a FileReader object in conjunction with the AudioContext decodeAudioData method. Reading and decoding takes place in the background, and success and failure are indicated by which of two callback functions are executed, as this code demonstrates:

```
// load a file from client
function openaudio()
{
  var file = document.getElementById('filechoice').files[0];
```

```
  var filename = file.name;

  var reader = new FileReader();

  reader.onload = function {
       var filedata = e.target.result;
       context.decodeAudioData(
            filedata,
            function onSuccess(buffer) {
                 signal = new Array(buffer.length);
                 var srcbuf = buffer.getChannelData(0);
                 for (i=0;i<buffer.length;i++) signal[i] =
srcbuf[i];
                 },
                 function onFailure() {
                      trace("decodeAudioData failed on
"+filename);
                 }
            );
  };

  reader.readAsArrayBuffer(file);
}
```

To play an audio signal, we create a processing chain using AudioContext methods, then run the chain through once. The createBufferSource method creates an element in the chain where we can inject audio samples. We create a buffer to hold our signal and pass it to the BufferSource. We then connect the BufferSource object to the output channel (context.destination), and kick off replay with its start() method.

```
// play some audio
var    sendsrc;
function playaudio(sig)
{
  var nsamp = sig.length;

  // create audio buffer source node
  sendsrc = context.createBufferSource();
  sendbuf = context.createBuffer(1,nsamp,context.sampleRate);

  // copy in the signal
  senddat = sendbuf.getChannelData(0);
  for (i=0;i<nsamp;i++) senddat[i] = sig[i];

  // kick it off
  sendsrc.buffer = sendbuf;
  sendsrc.loop = false;
  sendsrc.connect(context.destination);
  sendsrc.start(context.currentTime);
}
```

To stop the audio playing, it is possible to call the BufferSource stop method:

```
sendsrc.stop()
```

4 Advanced web audio functionality

In this section we highlight additional JavaScript objects and functions available through the web audio API which allow us to load audio from the server, to save audio to the client machine, to record audio and to process audio signals.

To load an audio file from the server, the XMLHttpRequest object can be used to transfer the file to the browser, then the decodeAudioData method of the AudioContext object is used to create an array of sample values. In the code below, note how the loading of the file is conducted in the background and the loadaudio functions returns before the data is actually available.

```
// load an audio file from server
var signal=[];
function loadaudio(aname)
{
   // Note: this loads asynchronously
   var request = new XMLHttpRequest();
   request.open("GET", aname, true);
   request.responseType = "arraybuffer";

   // callback loads signal into global buffer
   request.onload = function() {
        context.decodeAudioData(
             request.response,
             function onSuccess(buffer) {
                  signal = new Array(buffer.length);
                  var srcbuf = buffer.getChannelData(0);
                  for (var i=0;i<buffer.length;i++) signal[i] =
srcbuf[i];
             },
             function onFailure() {
                       alert("decodeAudioData failed");
             }
        );
   };

   // get file
   request.send();
}
```

Samples are stored as floats in the range -1.0 to +1.0 and converted to the AudioContext sampling rate. The decodeAudioData method supports a number of audio file formats, including MP3.

To save a signal back to the client machine, we create a WAV file in memory then trigger a download request by faking a click to a hyperlink. We use methods of a DataView object to gain access to a byte buffer and write a 16-bit version of the audio signal to the buffer complete with a WAV file header:

```
// set bytes in a buffer
function writeUTFBytes(view, offset, string)
{
  var lng = string.length;
  for (var i = 0; i < lng; i++) {
       view.setUint8(offset + i, string.charCodeAt(i));
  }
}

// make a WAV file from signal (16-bit mono)
function makeWAV(signal)
{
  var buffer = new ArrayBuffer(44 + signal.length * 2);
  var view = new DataView(buffer);

  // RIFF chunk descriptor
  writeUTFBytes(view, 0, 'RIFF');
```

```
  view.setUint32(4, 44 + signal.length * 2, true);
  writeUTFBytes(view, 8, 'WAVE');
  // FMT sub-chunk
  writeUTFBytes(view, 12, 'fmt ');
  view.setUint32(16, 16, true);
  view.setUint16(20, 1, true);
  view.setUint16(22, 1, true);
  view.setUint32(24, context.sampleRate, true);
  view.setUint32(28, context.sampleRate * 2, true);
  view.setUint16(32, 2, true);
  view.setUint16(34, 16, true);
  // data sub-chunk
  writeUTFBytes(view, 36, 'data');
  view.setUint32(40, signal.length * 2, true);

  // write the PCM samples
  var lng = signal.length;
  var index = 44;
  for (var i = 0; i < lng; i++) {
       view.setInt16(index, signal[i] * 30000, true);
       index += 2;
  }

  // make final binary blob
  var blob = new Blob ( [ view ], { type : 'audio/wav' } );
  return blob;
}

// save file
function saveaudio(sig)
{
  // create a hyperlink and fake a mouse click on it
  var a = document.createElement('a');
  a.href = window.URL.createObjectURL(makeWAV(sig));
  a.download = 'download.wav';
  var event = document.createEvent("MouseEvents");
  event.initMouseEvent(
       "click", true, false, window, 0, 0, 0, 0, 0,
       false, false, false, false, 0, null
  );
  a.dispatchEvent(event);
}
```

To make a recording using the microphone on the client machine, we first make use of the navigator.getUserMedia method to gain access to the microphone, then we use the AudioContext object set up a processing chain from the microphone to a script which siphons off the data passing through it into a global buffer. For security reasons, the getUserMedia function pops up a dialog to the user requesting confirmation that the script may access the microphone.

```
// start audio processing
var micsource=null;
var capturenode=null;
var recording=0;
function startrecording(stream)
{
  // create the microphone source
  micsource = context.createMediaStreamSource(stream);

  // create a processing node to capture the data
  capturenode = context.createScriptProcessor(8192, 1, 1);
  capturenode.onaudioprocess = function(e) {
```

```
        if (recording) {
             // only save data if recording flag is set
             var buf=e.inputBuffer.getChannelData(0);
             for (i=0;i<buf.length;i++) signal.push(buf[i]);
        }
  };

  // connect microphone to processing node and to output.
  micsource.connect(capturenode);
  capturenode.connect(context.destination);
}

// start/pause recording
function recordpause()
{
  // restart acquisition after pause
  if (!recording) {
        signal = new Array();
  }

  // first time only request use of microphone
  if (micsource==null) {
        // accommodate different names in different browsers
        navigator.getMedia = ( navigator.getUserMedia ||
                          navigator.webkitGetUserMedia ||
                          navigator.mozGetUserMedia ||
                          navigator.msGetUserMedia);
        navigator.getMedia(
             {audio:true},
             startrecording,
             function() { alert('getUserMedia() failed'); }
        );
  }

  // start/pause function
  recording = 1 - recording;
}
```

The first time the recordpause function is called, the recorded signal buffer is reset and the microphone is acquired. The second time, the recording is paused. In this code the recording is never actually stopped, merely halted from adding to the captured signal. This means that recording may be restarted without re-acquiring the microphone which would have caused another screen confirmation.

To demonstrate how some signal processing might be applied to a signal, we implement below a non-recursive low-pass filter at 1000Hz using the window method, then apply it to the audio signal through convolution:

```
// sinc function sinc(x) = sin(x) / x
function sinc(x)
{
     return Math.abs(x)<1.0E-10 ? 1 : Math.sin(x)/x;
}

// build non-recursive low-pass filter.
function nrlowpass(freq,ncoeff)
{
     // create symmetric buffer
     var nhalf=Math.floor(ncoeff/2);
     var filt=new Float32Array(2*nhalf+1);
     // calculate sinc function
     var omega=2*Math.PI*freq;
```

```
     for (var i=0;i<=nhalf;i++) {
          filt[nhalf+i]=filt[nhalf-i]=omega*sinc(i*omega)/Math.PI;
     }
     // Hamming window
     for (var i=0;i<=2*nhalf;i++) {
          filt[i] = filt[i] * (0.54-0.46*Math.cos(i*Math.PI/nhalf));
     }
     return filt;
}

// apply a filter to audio
function filteraudio()
{
     var lpfilt=nrlowpass(1000/context.sampleRate,31);
     var fsignal=new Float32Array(signal.length);
     // convolution
     for (var i=0;i<signal.length;i++) {
          var sum=0;
          for (var j=0;j<lpfilt.length;j++) {
               if ((i-j)>=0) sum += signal[i-j]*lpfilt[j];
          }
          fsignal[i]=sum;
     }
     signal=fsignal;
}
```

Other aspects of JavaScript programming are important for building software analysis tools, but are outside the scope of this article. In particular, worker threads are useful mechanisms for performing long calculations in the background without tying up the user interface; and the window.requestAnimationFrame() function is useful in building animations which synchronize to the display refresh rate.

5 Web audio software development

It is not always necessary to program web applications from scratch, since there are an increasing number of freely available libraries of standard functions to reduce development time. Perhaps the most well-known is JQuery[3], but we mention a few libraries directly relevant to speech analysis below.

5.1 Graphing Libraries

Web applications can create graphical elements as well as text. Modern web browsers support both pixel-based and vector-based drawing in 2 and 3 dimensions. For speech signal analysis applications, a common requirement is to produce mathematical graphs and charts, and a library of graph drawing functions provides a simple means for creating graphs without the need to build them from primitives such as lines and dots.

The Flotr2 graph plotting library[4] is a set of JavaScript objects and functions for plotting simple data plots and charts. It is open source and free to use. The Flotr2 library supports all major browsers including mobile, and can produce scatter plots, line plots, bar plots and pie charts.

[3] https://jquery.com/

[4] http://www.humblesoftware.com/flotr2/

The Highcharts graph plotting library[5] is another pure JavaScript library for plotting graphs. It has more options than Flotr2 and is a little more complex to use. Highcharts is a commercial product, but is free for personal use. The Highcharts library supports all major browsers including mobile, and can produce scatter plots, line plots, bar plots, pie charts, boxplots and many specialised plots.

5.2 Mathematical Libraries

Although JavaScript comes with a standard set of mathematical functions, it is often useful to be able to call on existing libraries of mathematical functions that support signal processing or statistics.

DSP.js is a comprehensive digital signal processing library for JavaScript[6]. It includes many functions for signal analysis and generation, including Oscillators (sine, saw, square, triangle), Window functions (Hann, Hamming, etc), Envelopes (ADSR), IIR Filters (lowpass, highpass, bandpass, notch), FFT and DFT transforms, Delays and Reverb.

SimpleStatistics is a library of basic statistical functions[7] for performing descriptive and inferential statistics, including regression.

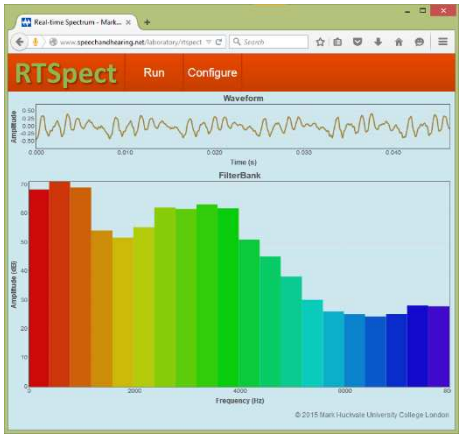
5.2.1 Examples

The following example web applications were written by the author and chosen to demonstrate the functionality that can be achieved using only HTML, CSS and JavaScript within a web browser.

RTSPECT

www.speechandhearing.net/laboratory/rtspect

RTSpect provides a real-time spectrum display from the user's microphone with waveform, spectrum and filterbank graphs. The application implements a real-time discrete fourier transform and performs graphical animation using the Flotr2 library.



[5] http://www.highcharts.com/

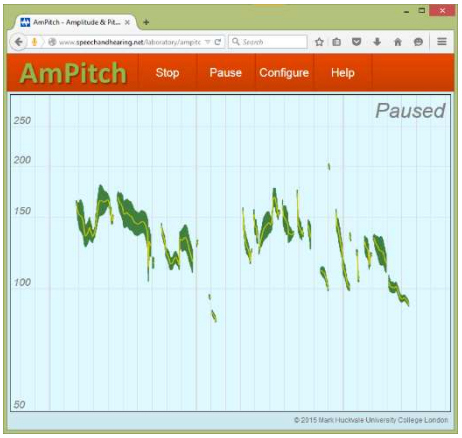
[6] https://github.com/corbanbrook/dsp.js/

[7] http://simplestatistics.org/

AMPITCH

www.speechandhearing.net/laboratory/ampitch

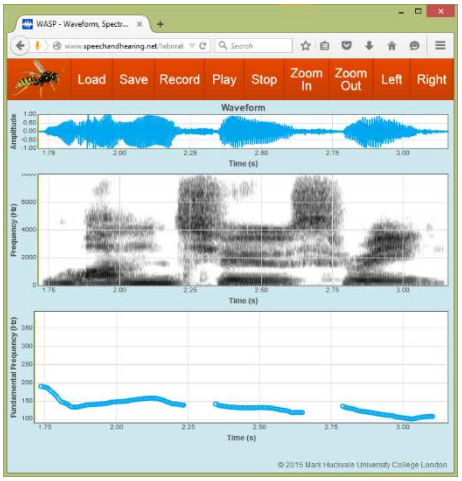
AmPitch provides a real-time amplitude and pitch track display from the user's microphone. The application implements an autocorrelation based fundamental frequency estimation algorithm and scrolling animation using the JavaScript animation methods.



WASP

www.speechandhearing.net/laboratory/wasp

WASP allows the user to record speech from the microphone and to display its waveform, spectrogram and pitch track. The application implements the SWIPE pitch estimator (Camacho & Harris, 2008) and spectrogram calculation. These run in worker threads since neither work in real time on most devices.



IMPROS

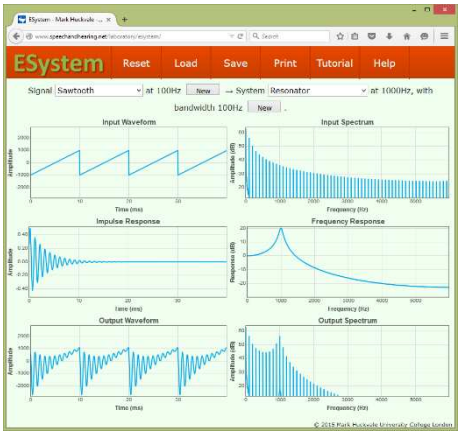
www.speechandhearing.net/laboratory/impros

ImPros is designed as a tool to improve the prosody of language learners. The user can record a sentence and compare its prosody with a teacher's version. The application implements mel-frequency cepstral coefficient (MFCC) calculation together with the SWIPE pitch estimation algorithm and dynamic programming time alignment.

ESYSTEM

www.speechandhearing.net/laboratory/esystem

ESystem is a tool for teaching and learning signal and systems theory. The application implements a general-purpose filtering library and fourier analysis. It uses the Flotr2 graph library.



6 Discussion and the future

Tablet computers may never fully replace conventional laptop and desktop computers for some applications. But their increasing number, power and ubiquity mean that software developers cannot shy away from making their tools and applications available on these platforms. This article has shown that at least some speech analysis tools originally developed for the Windows platform can be made to run fairly well as web applications within the browser on tablets thanks to the web audio API.

Some incompatibilities between computing platforms remain, particularly in the area of the web audio API which is still quite new. Apple iOS seems to put more constraints on how the AudioContext object is used compared to Android. These problems will be overcome in time, and the future will surely see more web audio applications like the ones described in this article.

In the future, there is scope for more sophisticated use of the web application environment, particularly through the exploitation of cloud computing and social networking. The interconnectedness of tablet computing allows for new kinds of collaborative work in which data may be collected and analyzed, and the results shared. We are beginning to see applications for the collaborative construction and labelling of speech corpora, the exploitation of native language speakers across the globe for phonetic analysis and pronunciation training, or the running of experiments in production and perception with hundreds of subjects on their own phones.

The open nature of web programming could be exploited to help advance the field of speech tools if authors are willing to share implementations of state-of-the-art algorithms within the web application framework. I am hopeful that libraries of speech analysis algorithms will be made available in the same way as the graphics and mathematical libraries mentioned above.

References

Huckvale, M. A., Brookes, D. M., Dworkin, L. T., Johnson, M. E., Pearce, D. J., Whitaker, L. (1987). The SPAR Speech Filing System. *European Conference on Speech Technology*, Edinburgh, 1987.

Carmacho, J. G. H. (2008). A sawtooth waveform inspired pitch estimator for speech and music. Journal of the Acoustical Society of America, 124, 1638-1652.

BOOK REVIEWS

Szypra-Kozłowska, Jolanta (2014):
Pronunciation in EFL Instruction.
Second Language Acquisition Series.
Bristol, UK: Multilingual Matters. 249 pp.,
Price: Hardback: ISBN 9781783092611, U.S. $ 159.95,
Paperback: ISBN 9781783092604, U.S. $ 49.95, £24.95, €29.95

Reviewed by: **Chantal Paboudjian**
University of Provence, Aix-en-Provence, France
e-mail: ChPaboudjian@aol.com

This publication, recommended by John Wells and John Maidment, is authored by **Jolanta Szpyra-Kozłowska,** an Associate Professor of English Linguistics and Chair of the Phonetics and Phonology Unit in the Department of English at Maria Curie-Skłodowska University, Lublin, Poland. Jolanta Szpyra-Kozłowska has already published 7 books and over 100 papers on English and Polish phonology, the phonology-morphology interaction, the acquisition of English phonetics and phonology by Poles, pronunciation pedagogy, foreign accent perception and gender linguistics.

This book addresses the issue of selection of pronunciation models for EFL learners. It is a relevant question, as the number of learners of English as a Foreign Language (EFL) is now estimated at around 1.5 billion, and as the concept of *English as an International Language* (EIL) or *English as a Lingua Franca* (ELF) has greater impact.

The book first presents arguments for the importance of pronunciation and provides an overview of recent debates about the choice of pronunciation models (Jenkins, 2000; Kachru, 1986; Lewis, 2005; Setter, 2010; Walker, 2011). It then underlines that pronunciation theory and assumptions, such as the supremacy of some models, the primacy of suprasegmentals, and the need for native instructors, have been challenged. It therefore deals with practical aspects of phonetic instruction while providing answers to a series of questions *English Foreign Language* (EFL) teachers are facing.

The volume contains four chapters with a significant number of endnotes, a bibliography, as well as subject and author indexes. Chapter 1 asks what pronunciation should be taught to foreign learners of English, Chapter 2 establishes pronunciation priorities for EFL learners and Chapter 3 deals with the issue of effective phonetic instruction while proposing a holistic multimodal approach. Chapter 4 sums up the main points discussed in the first three chapters.

As the book tries to be both general and specific, each chapter is divided into two parts (Part A and Part B). Part A presents a general theoretical discussion and Part B

verifies the theoretical claims raised in Part A through evidence provided by studies with Polish learners of English.

Chapter 1. English Pronunciation Teaching: Global versus Local Contexts (pp. 1-66) first shows the need to teach and learn the pronunciation of a foreign language, an important, yet often neglected, aspect of language. It then focuses on the goals of pronunciation teaching and learning and on the selection of an appropriate English pronunciation model. After a critical evaluation of the EFL and the EIL (or ELF) approaches, a concept carrying the potential of reconciling the two opposing views, a *Native English as a Lingua Franca* (NELF), is introduced and comparisons of the features of the three approaches are summarized in a table. A few pages are also dedicated to the distinction between EFL and *English as a Second Language* (ESL). Factors for diagnosing the local educational context of EFL instruction and learner-related and teacher-related determinants of pronunciation instruction are reported. The educational context factors include the national language policy, teacher preparation or the curriculum, while the learner- and teacher-related factors include students' goals, expectations and motivations, as well as and teachers' experience, involvement or attitude as to the role of pronunciation.

Chapter 2. Global and Local Pronunciation Priorities (pp. 67-139) deals with the quest to identify the factors which are most relevant for achieving intelligible pronunciation. Part A discusses the major aspects of pronunciation theories, then describes and evaluates recent proposals on pronunciation theories (Collins and Mees' Pronunciation Error Ranking, 2003; Cruttenden's Amalagam English and International English, 2008; and Jenkins' Lingua Franca, 2000). Based on these proposals, the author suggests that EFL phonodidactics instructors should focus on words usually mispronounced by learners and words which prevent intelligibility more than sounds and prosodic patterns. The nature of such words and the reasons for their mispronunciation are considered, as well as the impact of written forms on pronunciation. Finally, the chapter addresses the segmentals vs. suprasegmentals debate, i.e., should speech sounds or prosodies be the priority for EFL learners? The author suggests that the phonetic distance between L1 and L2 be considered first and that the impact of specific segmental and prosodic deviations from the L2 on intelligibility be evaluated by empirical research. Part B mentions studies supporting the claims made in Part A concerning phonologically deviant words which hinder the intelligibility of phonetically difficult words and the source of their mispronunciation. A study on identifying pronunciation priorities for Polish learners is also summarized.

Chapter 3. Pronunciation Inside and Outside the Classroom: A Holistic Multimodal Approach (pp. 140-224) is devoted to aspects of instructional procedures in and outside the language class. As those need to be both effective and attractive, as well as learner- and teacher-friendly, the author first discusses the importance of developing learners' concerns for good pronunciation and proposes a holistic motor-cognitive-multimodal approach to successful phonetic instruction. This approach has four main components: (1) articulatory training, (2) auditory training, (3) explicit phonetic and phonological procedure instructions to understand how L1

and L2 sound systems work and how the interferences between them affects performance, and (4) use of multisensory reinforcements in the acquisition of L2 phonetics and the presentation techniques. The necessity of pronunciation learning outside the classroom, the development of students' autonomy, the need to provide feedback of learners' phonetic performance and the correction of errors are further described. Part B summarises empirical studies on the effectiveness and attractiveness of pronunciation teachings, such as the use of phonetic drills, articulatory description, phonemic transcription, comparison of L1 and L2 phonetic systems, songs, poems and drama elements, and error correction to verify the claims made in Part A.

Chapter 4. Concluding Remarks (pp. 225-233). The 10 pages of this chapter briefly sum up the major claims made in the preceding chapters and indicate those areas of modern English phonodidactics for foreign learners that need closer attention from EFL pronunciation instructors. Thus, a series of 30 issues, which must be considered prior to teaching, are summarized.

This publication considers a fundamental and practical issue: the teaching of the pronunciation of a foreign language, which is the major aspect of a language for learners at a time when English has become a Lingua Franca. Its overviews of recent theories on the subject may also catch the reader's attention. Moreover, Szpyra-Kozłowska provides convincing arguments resting on a research-based approach. In conclusion, we could say that although some theoretical paragraphs may appear daunting to some, the issues themselves are engaging. There should be two possible groups of readers for this book: English teachers - particularly those with an enthusiasm for pronunciation instruction and phoneticians, postgraduate researchers, as well as students of English and ELF theoreticians who should not remain indifferent to the theoretical and thought-provoking insights it contains.

References

Collins, B. S., & Mees, I. M. (2003). *Practical Phonetics and Phonology: A Resource Book for Students*. English Language Introductions. London. Routledge.

Cruttenden, A. (2008). *Gimson's Pronunciation of English*. Seventh edition, London: Hodder Education. Associated material at the Companion Website: www.hodderplus.co.uk/linguistics.

Jenkins, J. (2000). *The Phonology of English as an International Language.* Oxford: Oxford University Press.

Kachru, B. B. (1986) *The Alchemy of English: The Spread, Functions, and Models of Non-Native Englishes*. Oxford [Oxfordshire]: Pergamon Institute of English.

Levis, J. (2005). *Changing Contexts and Shifting Paradigms in Pronunciation. Teaching*. TESOL Quarterly, 39(3), 369–377.

Setter, J. (2008). Theories and approaches in English Pronunciation. In Monroy R., & Sanchez, A. (Eds.): *Theories and Approaches in English Pronunciation*. Universidad de Murcia: Servicio de Publacaciaones. 447-457.

Walker, R. (2011). *Teaching the Pronunciation of English as a Lingua Franca*. Oxford: Oxford University Press.

Sun-Ah Jun (Ed.) (2014):
Prosodic Typology: The Phonology of Intonation and Phrasing.
Oxford: Oxford University Press
ix + 462 pp + a CD., including: Preface, List of Contributors and Index,
wav files demos,
ISBN 019-924963-6, price: £60.00**,** Hardback

Reviewed by: **Vered Silber-Varod**
The Research Center for Innoavtion in Learning Technologies,
The Open University of Israel, Ra'anana, Israel
e-mail: vereds@openu.ac.il

Sun-Ah Jun is a Professor in the Department of Linguistics at the University of California, Los Angeles (UCLA). Formerly a student of Mary E. Beckman, Jun has worked on the phonetics and phonology of Korean prosody, published in the book *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure* (Garland Publishing, Inc., 1996). She is currently the President of the International Circle of Korean Linguistics (2014-2016).

This edited volume is the second volume of Jun's *Prosodic Typology: The Phonology of Intonation and Phrasing* (OUP, 2005). Like the first volume, *Prosodic Typology II*, is based on a single basic theoretical framework: The Autosegmental-Metrical (AM) model of intonational phonology (inter alia, Beckman and Pierrehumbert 1986). Unlike the first volume, half of the languages, which vary in their word prosody as well as their geographic distribution, are understudied languages or researched through fieldwork.[8]

Prosodic Typology II is organized in seventeen chapters, beginning with an informative Introduction (chapter 1 by S.-A. Jun, pp. 1-5) and closing with two didactic chapters: A chapter that discusses the methodology of studying intonation: from data collection to data analysis (chapter 16 by S.-A. Jun and J. Fletcher, pp. 493-519), and a summary of the three key parameters in prosodic typology, as suggested by Sun-Ah Jun: prominence type, word prosody, and macro-rhythm (chapter 17 by S.-A. Jun, pp. 520-539).

The other fourteen chapters proceed as follows (in order of appearance): 2. Sónia Frota: The intonational phonology of European Portuguese (pp. 6-42). 3. Pilar Prieto: The intonational phonology of Catalan (pp. 43-80).4. Sameer ud Dowla Khan: The

[8] In the first volume, Jun edited descriptions of the intonation and the prosodic structure of thirteen languages: German, Greek, Dutch, Serbo-Croatian, Japanese, Korean, (Pan-) Mandarin, Cantonese, Chickasaw (a Native American Indian language), Bininj Gun-wok (an Australian aborigine language, AKA Mayali), varieties of Italian, Four Dialects of English, and Swedish.

intonational phonology of Bangladeshi Standard Bengali (pp. 81-117). 5. Elinor Keane: The intonational phonology of Tamil (pp. 118-153). 6. Chad Vicenik and Sun-Ah Jun: An autosegmental-metrical analysis of Georgian intonation (pp. 154-186). 7. Anastasia M. Karlsson: The intonational phonology of Mongolian (pp. 187-215). 8. Anja Arnhold: Prosodic structure and focus realization in West Greenlandic (pp. 216-251). 9. Janet Fletcher: Intonation and prosody in Dalabon (pp. 252-272). 10. Shelome Gooden: Aspects of the intonational phonology of Jamaican Creole (pp. 273-301). 11. Bert Remijsen, Farienne Martis, and Ronald Severing: The marked accentuation pattern of Curaçao Papiamentu (302-323). 12. Carlos Gussenhoven: Complex intonation near the tonal isogloss in the Netherlands (pp. 324-364). 13. Dana Chahal and Sam Hellmuth: The intonation of Lebanese and Egyptian Arabic (pp. 365-404). 14. Gorka Elordieta and José Hualde: Intonation in Basque (pp. 405-463).15. Yoshuke Igarashi: Typology of intonational phrasing in Japanese dialects (pp. 464-492). The organization of these chapters is well justified in the Introduction (pp. 3-4), yet, each of these fourteen chapters stands alone, and therefore the chapters can be read in any order, so that the reader can choose the language that most intrigues her/him.

In addition to the book, the volume has its own website, where interested readers can listen to (close to) 400 sound files associated with all the figures included in the book and that exemplify prosodic phenomena discussed in the chapters (www.oup.co.uk/companion/jun2).

Since this volume deals with fourteen languages, and in few cases, with group of languages, differences in the descriptions are expected, yet all chapters present the same topics: The word prosody of the language, methods of data collection, the tonal inventory, intonational characteristics of focus prosody, and the prosodic structure of the language. All chapters (except for Japanese (chapter 15)), also present intonation of various sentence types. The descriptions of the tonal categories and prosodic patterns for each language and dialect is carried out by the tonal labeling system of ToBI (Tones and Break Indices). This unified annotation system, together with the coherent terminology, symbols, and conventions is what enables a clear prosodic comparison across languages. Due to the space limitations of this review, in the following, I will highlight some core findings above the word level prosody, of each of the fourteen studies.

Chapter 2 European Portuguese (EP): This chapter describes the three domains of prosodic phrasing in EP: Prosodic Word (PW), which is the domain of many segmental and prominence properties; Phonological Phrase (PhP), which only plays role in rhythmic and prominence related phenomena; and Intonational Phrase (IP), which is the main domain of prosodic manifestations (segmental, durational, tonal and of prominence). The most salient feature of EP is "the sparseness of pitch accents with the IP." (p. 40). An interesting interface between segmental processes and intonational structure of questions is presented – EP does not compress or truncate tonal structures (p. 26): When a sequence of tones is aligned with a single syllable, the segmental string is extended either by lengthening of the nuclear vowel or by a vowel epenthesis.

Chapter 3 Catalan: This chapter describes the properties of three domains of prosodic phrasing in Catalan: PhP, Intermediate Phrase (ip), and IP. Also, the description comprises a single type boundary tone for both ip and IP. The chapter mainly discusses the speech acts and pragmatic context of intonation. The author points out several issues that are still unresolved with respect to Catalan, concerning the AM framework. Among them is the contrastive use of tonal alignment that was found in the rising pitch accents. The author concludes that "In general, the Catalan data provides evidence that the fixed tonal alignment in the standard definition of bitonal pitch accents does not correspond with the empirical data." (p. 79).

Chapter 4 Bengali: This chapter is based on recordings from a large number of speakers in varied contexts. The model of Bangladeshi Standard Bengali intonational phonology distinguishes the three domains: Accentual phrase (AP), ip, and IP. Moreover, the chapter presents the rich tonal inventory and phonological interactions between tones, with exceptional attributes of the focus high tone.

Chapter 5 Tamil: This chapter show how the inventory of pitch accents in Tamil is highly restricted, even when manipulating the focus type (broad vs. narrow). Moreover, the chapter distinguishes two prosodic domains: AP and IP, while ip structure requires further investigation. Declarative and interrogative tonal structures are also inspected.

Chapter 6 Georgian: The authors propose that Georgian has three levels of phrasing: AP, ip, and IP. Each of these phrases has an inventory of three final boundary tones, which mark their right boundary. IPs and ips are also marked with phrase-final lengthening. APs are also marked with rich tonal marking: Four phrase-initial Pitch Accents; and one Phrase Accent, which is associated with antepenultimate syllable.

Chapter 7 Halh Mongolian: The proposed analysis of Halh Mongolian does not recognize any Pitch Accents in this language. Rather, Halh Mongolian seems to belong to edge-prominence (according to Jun's (2005) typology). Thus, pragmatic meanings are achieved within the domain of phrase final boundary tone.

Chapter 8 West Greenlandic: The author proposes that the prosodic hierarchy of West Greenlandic consists of three units: Mora, PW and IP. The Mora is the tone-bearing unit; The PW is marked by tonal contour and a single underlying word-final tone (HLH sequence), and the IP is the domain of declination.

Chapter 9 Dalabon: This chapter deals with an endangered Australian language – Dalabon. It is assumed that Dalabon is a stress accent language, but with edge-marking phrasal elements, in the realization of focus. The IP is the highest unit in the prosodic hierarchy, with an AP unit below it, then the PW and Foot.

Chapter 10 Jamaican Creole (JC): JC seems to have two phrasing units above the word: ip and IP. The falling pitch accent is the most prevalent (among the four accent categories in JC), and is realized in a variety of utterance types. In addition, there is an inventory of three right-edge boundary tones at the IP level, and two phrase accents at the ip level. JC also uses pitch accent to mark emphatic focus, which

together with the syntactic reorganization, is used as a double complex foci (prosodic and syntactic) process.

Chapter 11 Curaçao Papiamentu: The chapter deals with the interaction between lexical tone and intonational prominence of the Curaçao dialect of Papiamentu. This study found that the formal realization of the lexical tone is not constant across sentence types. Thus, both lexical and sentence-level information weigh in the specification of tones.

Chapter 12 Helden and Venlo Dutch: In this chapter, the author claim that the dialects of Helden and Venlo are tone languages, and because of that are distinctive from their perceptually close dialects of Standard Dutch. The chapter describes, in detail, the two IP-final declarative contours: for accent 1 (toneless) and accent 2 (includes tone); and the more complex patterns for IP-final interrogatives contours, for the two accents. The two dialects are also compared to the dialect of Roermond, and a phonetic and phonological dialect continuum is suggested.

Chapter 13 Lebanese Arabic and Egyptian Arabic: This chapter compares two intonational phonology systems of two different Arabic dialects: Lebanese Arabic and Egyptian Arabic. The authors have found similarities, as well as differences, in the inventory and distribution of pitch accents in these two dialects, but boundary tones, and to some extent, phrase tones are similar in both dialects. Since the study compares the findings and analysis of two separate research settings, the authors call for a future study, which will explore different Arabic varieties by means of directly parallel comparison.

Chapter 14 Basque: The chapter concentrates on two varieties of the Basque dialects: Northern Bizkaian Basque (NBB) and Standard Basque (SB). The two varieties are different in their word level prosody, as NBB has a lexical tone system (i.e., pitch accent system) and SB is a stress-accented dialect. Nevertheless, the analysis of both varieties reveals three prosodic units: AP, ip, and IP. Declarative, interrogative, imperative, and exclamative sentence types are also examined with respect to their basic intonational contours.

Chapter 15 Typology of intonational phrasing in Japanese dialects: The chapter is concerned with Japanese dialects without lexically-specified tone, and tries to find the intonational properties that distinguish between its two subgroups (one-pattern accent vs. accentless). The author proposes a classification of the two [-lexical tone] dialects solely based on a parameter concerning intonational phrasing ([±multiword AP]). He then summarizes that the variability observed in tonal shape can be derived by the two binary parameters [± lexical tone] and [±multiword AP], and does not serve as a classification property on its own. The second part of the chapter, further applies this classification to four other dialects of the [+lexical tone] dialectic group.

As mentioned above, the volume ends with two general chapters. The first is about the methodology of studying intonation, and the second outlines the prosodic typology encompassing the fourteen languages' specific descriptions, which is suggested as a unified typology for the Intonational Phonology framework.

Chapter 16 Methodology of Studying Intonation: From Data Collection to Data Analysis provides extremely valuable information on various methodological

aspects in the study of intonation. With its clear language and organized structure, the chapter systematically describes how to accomplish a state of the art intonational research program. The first phase is dedicated to the methodology for designing a prosodic database for each language type, in terms of lexical word structures, and explains how to build multi-word sentences in order "to tease apart the word prosody from phrasal prosody…" (p. 494). Then the methodology of collecting other sentence types than declaratives is introduced, followed by designing the prosody of focus. The next step, according to the authors, is to examine the prosodic constituents (i.e., units) of a certain language and to set the relevant prosodic hierarchy. The last step in the design of intonational phonology research is to decide on the speakers and the lab settings, as well as how to prepare the scripted prompts carefully. Following this design, the authors give Dos and Don'ts on how to collect the data, how to process the fieldwork and the recordings. The last section of the methodological chapter deals with the analysis. Here, the critical use of acoustic phonetic software is introduced. The chapter continues by presenting the annotation phase, where decisions on tonal categories and symbols are crucial. As a summary, the chapter presents all tonal categories and diacritics that can be used in describing and analyzing intonation contours (pp. 517-518). This is also the part which the author claims "should apply to all languages if analyzed in the framework of intonational phonology." (p. 519).

Chapter 17 Prosodic Typology: By Prominence Type, Word Prosody, and Macro-rhythm presents the revised model of prosodic typology [compared to the first volume (Jun, 2005)]. This model consists of three parameters: Type of prominence marking, word prosody, and macro-rhythm. According to Jun, this typology allows for comparison between languages that belong to the same type of word prosody and prominence marking, but differ by the global pattern of phrase-medial intonation contour. On the other hand, this model captures intonation patterns across languages that do not share the same type of word prosody and prominence marking. The innovative parameter of macro-rhythm, proposed in this book, is defined here as phrase-medial tonal rhythm, which captures the regularity of phrase-medial intonation patterns across languages.

Last, in addition to the core chapters, the book contains a List of Abbreviations (pp. xiii-xv); a 40 page Reference list and a 7 page Index of key terms in prosodic typology, which I found very useful (although the pages are not always precise, for example "secondary accent" was not found on page 401). This, together with close to 400 annotated and segmented pitch track figures, make this book a good reference and guide for researchers and graduate students working on intonation and prosody, especially for those who are willing to build or refine a phonological model of intonation in an under-studied language.

References

Jun, S. A. (Ed.) (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing.* Oxford University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology,* 3(1), 255-309.

Johanneke Caspers, Yiya Chen, Willemijn Heeren, Jos Pacilly, Niels O. Schiller, and Ellen van Zanten (Eds.) (2014):
***Above and Beyond the Segments*: *Experimental Linguistics and Phonetics*.**
Amsterdam/Philadelphia: John Benjamins Publishing Co.
363 pp., including a Foreword (pp. XI-XII) and Index (pp. 359-363),
ISBN-13: 978-9027212160, ISBN-10: 9027212163,
Hard back, Price: €105.00, US $ 158.00

Reviewed by: **Judith Rosenhouse**
SWANTECH Ltd., Haifa 3628412 Israel
e-mail: judith@swantech.co.il

This Festschrift was presented to Prof. Vincent van Heuven, Professor of Experimental Linguistics and Phonetics at Leiden University on his retirement in 2014. The volume contains 27 papers dealing with various languages and experimental phonetic topics, which were written by colleagues and PhD students of Prof. van Heuven. We can divide the papers into three clusters: 1. languages, 2. topics, and 3. analysis methods.

1. The languages dealt with in the book are (mainly) Dutch, North-West Indo-Aryan (spoken in the northern part of India, Pakistan and the region), Zulu, English, Indonesian and Austronesian languages, Danish and Swedish, Greek, Chinese and Wenzhou Chinese (Chongqink dialects), Carib (the Cornelis Kondre dialect), Italian (the Sienese dialect of Tuscany), Tundra Yukaghir, and Agreer Dinka (spoken in South Sudan). Dutch, discussed in several papers, and English are studied also in comparison with some of the other languages.

2. Diverse phonetic topics are presented, including the following linguistic areas: phonetics-related subjects (acoustic phonetics, syllable monitoring, pauses, tonal coarticulation, laryngeals in Dutch, affricates in English and vowel duration categories as well as phrasal stress), prosody and intonation (effects of prosodic structure, tone, boundary tones, stress, pitch accent placement, in general and in L1 and L2, intonation, and tone in whispered speech), psycholinguistics (pauses, speech errors, perception of fricative devoicing), Sociolinguistics (age, level of education), dialect features (in Dutch and Chinese), Grammar, Semantics and etymological aspects in questions, and graphemic systems and their effect on the development of different languages.

3. The analysis in these chapters demonstrates the wealth of methods currently used in experimental phonetics. The papers begin with a literature survey, which is often rather detailed. The papers use methods, which confirm past assumptions or features, as well as enhance new approaches and theories. The papers include computational programs, databases, and various other ways of studying speakers'- or listeners' performance. Thus, many of the papers present figures and tables of their findings. Below is a summary of all the papers in the order of their appearance in the book:

Joan L. H. Baart, Tone and stress in North-West Indo-Aryan: A survey (1-13). The paper deals with the contrastive lexical tone of the languages in the title. These languages (a few hundred of them) are classified into three groups: Punjabi (Hindu-Urdu), Shinna, and Kalami. The analysis reflects the fact that several of them lost their breathy-voiced speech sounds and they lack tone-dependent words, others have tone and breathy voiced phonemes, and still others do not have tones, but have breathy-voiced phonemes. The tones can be complex, rising and falling, or simple tones. In some of these languages, word accent depends on syllable structure, changing when various grammatical suffixes are attached to the word. In other languages (e.g., Punjabi), word meanings change according to the tone type and its place. There are also languages where tone is historical, i.e., it is lost on the surface, but affects word structure. The variations are numerous, and not always expected. The authors end with a call for more study on these little-researched languages.

Tina Cambier-Langeveld, Maya van Rossum and Jos Vermeulen, Whose voice is that? Challenges in forensic phonetics (14-27). These authors mention in the beginning several difficulties in the field of Forensic Speech Recognition (FSR), mainly speaker variability, voice definition and terminological description and conclusions. In phonetics, voice quality is not always relevant for the research, but voice quality is very important in speaker identification, because it is "woven into the fabric of speech" (p. 17, quoting Laver, 1994: 2). These authors think that voice quality problems are due to the fact that it is considered componential, and mention Kreiman and Sidtis (2011) who present many examples against the assumption that voice quality can be described by feature lists. Kreiman and Sidtis (ibid.) suggest that its description is based on pattern recognition and feature analysis. The authors of this paper add that speech perception depends not only on the speaker, but also on the listener. It is therefore clear that this subject is a serious challenge for research. Cambier-Langeveld et al. suggest creating a "blind group" that will enable collection of similar elements in different voice examples (of the suspect and the original recording) and then checking for the similarities between them. The advantage of this method is that is does not oblige the analyzing phonetician to express a clear opinion about the recorded voices, but to base the report only on acoustic facts.

Johanneke Caspers, Pitch accent placement in Dutch as a second language: An exploratory investigation (28-41). This author studies the question of how pitch accent differs in native speakers from speakers of other languages, when they speak the same language. This paper compares Dutch-L2 speech of Polish, French, Chinese and Hungarian speakers with Dutch-L1 speakers. Checking accentability in Dutch words spoken by these examinees, revealed that they produced the necessary accents and most of the incorrect accents did not occur (i.e., they spoke mainly "correctly"). The Dutch-L2 speakers accentuated about 2/3 of the potentially accentible syllables in the test words, whereas the Dutch L1 accentuated them less (at about 1/3 of the options). But other differences between the language groups lead the researcher to study the differences between "plastic" and "non-plastic" languages (Dutch being plastic,

whereas French is not). Although the author notes that this study has its limitations, its findings reveal a strong effect of L1 (matching previous studies).

Lisa Lai-Shen Cheng and Laura J. Downing, The problems of adverbs in Zulu (42-59). After previous publications about the Zulu language by these authors, this paper begins with noting that in this language, adverbs can appear as prosodically-linked to the main sentence (utterance) or separated from it. Adverbs can appear in neutral and focused contexts, they can be single or multiple in a sentence, and certain syntactic structures affect joining or dis-joining the adverb to the sentence structure. These researchers designed a syntactic analysis tree, which includes the adverbs and shows this difference. (These points are demonstrated by examples, of course.) The authors finally suggest that Zulu adverbs have a nominal nature and therefore can be selected by verbs.

Crit Cremers and Maarten Hijzelendoorn, Meaningful grammar is binary, local, anti-symmetric, recursive, and incomplete (60-70). This paper deals with the calculation of linguistic meaning, based on their program Delilah (designed for this purpose) and an improved program named “incomplete”. From the work on these programs, the researchers draw the conclusion that linguistic meaning includes the five elements mentioned in the title of their paper. While describing the program and the necessary parameters for it, they analyze why these five elements are important for sentence comprehension by a computer program. Phonetics is an aspect of grammar, but it does not create correct word order. Phonetics, they write, reflects language as it is, whereas grammar reflects a language as it is never possible to be. Finally, they consider the five elements as more or less a good guess of grammar when it does what it has to do: pack and provide dynamic information about the language.

Anne Cutler and James M. McQueen, How prosody is both mandatory and optional (71-82). This paper begins with quoting Lehiste's (1970) and Bolinger's (1964) views on prosody. The former claims that prosody is a necessary part of any utterance, whereas the latter believes that one could understand a language even without intonation (e.g., in silent reading or monotonous speaking). Studies of Dutch and English reveal opposite phenomena: a stressed syllable in a word did not change its identification (in English), but did affect its identification in Dutch. The authors describe experiments with word accent stress and sentence stress (focus), which show different results in different conditions. The authors' conclusion is that both Lehiste and Bolinger are right. Prosody is necessary and exists in speech; and users/speakers heed prosody when it contributes to a distinctive meaning element in utterances, but can ignore it when it is not relevant for them.

Rob Goedemans and Ellen van Zanten, No stress typology (83-95). This paper deals with the Prosody of Indonesian Languages (PI), in the framework of that project, which formally began in 1957. At the end of that project, 500 languages were found in the database (StressTyp) and the knowledge accumulated there was published in many papers and books. That database and another one (Bailey, 1995) was merged in SPD (Stress Pattern Databases), a unified database (Heinz 2007). Goedemans and van Zanten created another database for the internet (StressTyp2), which was almost completed when this paper was written and is going to include 700 languages when

finished. These authors describe a few prosodic features of these no-stress dialects, and call for more studies based on new data from new languages.

Charlotte Gooskens and Renée van Bezooijen, The effect of pause insertion on the intelligibility of Danish among Swedes (96-108). Swedish and Danish are similar languages in many respects, but they differ in the area of accent. Swedish has a distinctive stress, which Danish does not have. Studies show that the acquisition of Danish is much slower than Swedish, and learners of Danish as an L2 find difficulties in acquiring it. The issue discussed in this paper is whether this difficulty depends on lack of distinctiveness in Danish. They tested it by adding pauses to 15 equal length sentences, compared to the same sentences, without pauses. The testing conditions differed to some extent from previous studies. The findings are however generally similar: most of the sentences with pauses were better understood than those without pauses. A deeper analysis found small, but non-significant additional differences. The main conclusion is that pauses before prosodic boundaries help sentence understanding (probably due to additional processing time for listeners).

Stella Gryllia, Intonation, bias and Greek NPIs: A perception experiment (109-119). This paper examines differences between three types of Greek interrogative sentences: (1) a negative sentence with an unaccented negation particle, (2) positive with an un-accented negation particle. 3. a negative sentence with an accented negation particle. The findings show a negative preference for type (1), a positive answer for type (2) and a less clear preference and negative answers for type 3. These findings match the research hypotheses and previous studies, but the innovation is that there is a bias for the positive answer in the case of a positive question when the interrogative particle is un-stressed.

Yan Gu and Aoju Chen, Information status and L2 prosody: A study of reference maintenance in Chinese learners of Dutch (120-130). The topic of this paper is information expression in Dutch, and the difference between given and accessiblility in Dutch. Mandarin-L1 Chinese learners of Dutch participated and were compared to a control group of Dutch native speakers while reading texts in these languages. The recordings were made in both the Netherlands and China. Word durations and pitch width were measured in both languages. The results from both languages show that the learners are influenced by pitch more than by duration. There were acquisition level-dependent differences between the average- and advanced-level learners as well.

Willemijn F. L. Heeren, Does boundary tone production in whispered speech depend on its bearer? Exploring a case of tonal crowding in whisper (131-143). Here we read about features of male and female whispered speech, which is a complicated subject. Studies found that although F0 was not available in whispered speech, listeners could distinguish interrogative sentences from statements in American English due to cues that differed from those used in normal speech. In this study, relative syllable duration, intensity, F1 – F3 frequencies, and normalized spectral energy in four frequency bands were examined. The findings matched previous studies. In general, vowel quality affected all findings, probably because of the vowel structure in the vocal tract: open articulation affected intensity more than closed

(vowel) articulation, and accent location affected boundary tone (rising or falling) perception. The author suggests that close prosodic events challenge production and perception in whispered speech.

Berend Hoff, The primacy of the weak in Carib prosody (144-151). Accents are usually associated with the stressed syllable in a word; but such a link does not exist in the Carib Cornelius Kondre dialect. Back in 1968, the author measured pitch differences in a stressed foot compared to the following feet by duration changes between first and second syllables of words. He found differences between iambic and trochaic patterns in these words in all the speakers, when the main stress was on the last syllable. In this paper, he compares the stress in 3- and 4- syllable words in iambic and trochaic patterns. After defining the stress rules in this dialect, he concludes that in any case, there is no connection between the prominence of a weak + long syllable and the place of the string stress. On the contrary, “the weaker kind of prominence comes first.” This state differs from the other two dialects studied by this author.

Jan H. Hulstijn and Sible Andringa, The effects of age and level of education on the ability of adult native speakers of Dutch to segment speech into words (152-164). The topic of this paper is people's age- and/or education-dependent ability to segment a speech passage into words. The starting point is the phenomenon that completion of speech-rate dependent elision of segments (consonants, vowels), syllables, and parts of whole syllables requires in young speakers less time than older speakers. This phenomenon (Basic Language Cognition, BLC) is studied here with 345 participants of various ages, native and non-native speakers of Dutch, and various educational levels. They had to count the number of recorded words they heard and write them down (no timing condition). The results revealed that in the writing task, the younger participants were faster and more accurate than the older ones. In the counting task, significant age and education effects were found. Participants with higher education, IQ and better working memory gave more correct answers than other participants. The education level effect had not been expected. These findings suggest more complex meta-cognitive activity than simply speech identification and classification.

Robert S. Kirsner, Doing grammatical semantics as if it were phonetics (165-173). Previous papers by this author (Kirsner, 1988, 1989) concerning Dutch “deze” and “dies” ('this') showed a semantic difference between these words in their range of relevance in the discourse. Here, the author presents statistically analyzed phonetic data of utterances with these words in the contexts of repetition and re-chunking. Another analysis refers to the intonation effect on two imperative forms in Dutch (imperative verb forms and infinitive forms, van Heuven, & Kirsner, 1999), statistically analyzed here for utterance-final pitch structures. The analysis finds that the verb form (imperative) has less of an effect (range) than the infinitive. Thus, the verb can be both real and pseudo-imperative, whereas the infinitive refers only to real imperatives.

Sara Lusini, Roberta D'Alessandro and Johan Rooryck, Phonetic aspects of polar questions in Sienese: An experimental approach (174-188). The “yes-no questions” in the dialect of Siena, Tuscany, in Italy, have two structures: (1) che fare and (2) che

+ verb (no “fare”). This paper analyzes 110 recorded sentences (simple and bi-clausal utterances) read by 11 speakers in questions with both types, with many acoustic differences between them, as the presented measurement data show. The main differences involve vowel duration before a pause (in the word “fare”) and pause duration. Other differences appeared in intensity, and pitch fall before the pause (if it existed). Pause did not appear in 27 sentences, but always appeared in the sentences without “fare”. Thus, sharp prosodic differences exist between these two structures, with timing differences as the main distinguishing feature.

Annecke Neijt, Etymological sub-lexicons constrain the graphemic solution space (189-202). This paper discusses the case of Dutch graphemics. Neijt analyses the Dutch writing system, in light of Neef's (2005) and van Heuven's (1994) approaches. Neef (2005) suggested a difference between indigenous words and foreign/borrowed words in German; van Heuven et al. (1994) suggested (for Dutch) several criteria, which enabled them to identify 90% of the words correctly, though errors also occurred. Neijt rejects Neef's approach and claims that Chomsky's “triple” approach (observation, description, explanation) in linguistic analysis should also be applied in non-systematic graphemics, due to the integration of words from many different languages in Dutch.

Sieb Nooteboom and Hugo Quené, Do speakers try to distract attention from their speech errors? The prosody of self-repairs (203-217). These authors study two manners of speech-error corrections: during speaking and after utterance completion. This topic, while not new, is very complex, as this paper demonstrates. Many parameters of erroneous vowels were measured in the speech of 38 Dutch speakers (e.g., timing of various utterance parts and errors, pitch and loudness maxima, minima, average, spectral slant, etc.). The authors found a strong correlation between phonetic features of the studied erroneous vowels and their manner of correction. For example, errors detected early were longer and had higher maximal and average loudness than errors detected later, for which lesser vocal effort was used. Nooteboom and Quené finally conclude that speakers have different strategies for correcting speech errors.

Cecile Odé, Field notes from a phonetician on Tundra Yukaghir orthography (218-229). This paper discusses problems of an endangered language with a small number of speakers, though published grammars and dictionaries exist and are used in media and at school, to some extent. Odé presents difficulties among different speakers of dialects of the same language, which are expressed in variable transcriptions of certain vowel- and consonant-phonemes. These differences make it difficult to acquire reading, writing, and comprehension of written texts, and these issues create difficulties for both school teachers and students. To solve these problems, Odé suggests using Kurilov's dictionary and transcription rules, despite the deviations from the rules found in his dictionary and the difficulties caused by this graphemic inconsistency.

Anne-France Pinget, Jans Van de Velde and René Kager, Cross-sectional differences in the perception of fricative devoicing (230-245). Dutch is known for the fact that /v/ is pronounced as /f/ in syllable onsets. This phenomenon varies, however, between Dutch dialects and therefore the paper compares this phenomenon in three

Dutch dialects. The experiment involved listeners who heard words from the three dialects and had to identify /v/ or /f/ in them. The data creation (nine grades, and their manipulation manner) and the findings are meticulously described, including listeners' response uniformity. Differences indicating gradual spread of this phenomenon from the north to the south appeared. The authors describe the three stages of this process (/v/ > /f/) and observe that periodicity is a very strong marker for it, whereas timing is not.

Bert Remijsen, Evidence for three-level vowel length in Ageer Dinka (246-260). This paper studies vowel durations in the Dinka dialect (spoken in the Western Nile region, in Sudan).In that language, vowel duration, tone and voice quality are separate distinctive features. Voice quality is modal (v) or fricative (f). There are four tones: High, Low, Fall and Rise / Mid. The author presents his method, which includes two structural models and his own hypothesis. The latter has four levels of morpho-lexical quantity in two linked levels (mid or end of the utterance), by four segmental systems and produced by speakers).This project had 360 items and the acoustical findings and their statistical analysis have verified that in this Agee dialect, there are three vowel duration contrasts which are related to morphology. The author finally notes that a similar system exists in the Luanyjang language.

Tony Rietveld and Niels O. Schiller, Phonetic accounts of timed responses in syllable monitoring experiments (261-274). This paper focuses on the manner of phonetic-acoustic level mapping onto the lexical level. Both phonemes and syllables can serve at the level of sub-lexical representation. The authors studied the interface between the phonetic-acoustic and lexical levels. They applied a monitoring technique, which requires participants to answer questions that deal with the crossover interaction between target and word (named the “syllable match effect”) within some time-frame. The studied consonant segments were /l/, /r/, /m and n/ and /k and p/ and the vowels /e:/, /o:/. The experiment did not confirm the hypothesis that there will be a triple interaction (consonant x studied word x target) and clear effects. However, listeners could not use the (phonetic) information about the identity of the pivot-consonant (in the target CVC) during the vowel articulation if the pivot- consonant was not affected by the preceding vowel (as happens when the consonant is a stop). A consonant like /r/ is affected by the preceding vowel and thus listeners can identify the consonant faster. CVC syllables (with /r/) reveal faster and more accurate identification than CV syllables. The assumption that acoustic information is applied during syllable examination is confirmed by the fact that Response Times for C1VC2 are in correlation with vowel duration when the pivot-consonant (C2) is an /l/, /r/ or nasal (but not a stop).

Franziska Scholz and Yiya Chen, The independence effects of prosodic structure and information status on tonal coarticulation: Evidence from Wenzhou Chinese (275-287). The paper investigates if and how much prosodic structures affect tone production in the Chinese Wenzhou tonal language. This issue of tonal coarticulation has been hardly studied for this language and for other tonal languages. The test material included four tones, in sentences with/without focus on the same lexeme (as in answers to yes-no questions), and the examined factors were left/right headword

position, structure (verb-adverb/verb-accusative noun), context (conflicting/non-conflicting) and focus (present/not present). The main effect was of context and position, but other significant interactions were also found. Thus, a general pattern was that tones in the rightmost position covered a greater portion of the speaker's vocal range and exhibited a magnified tonal movement across all conditions. Moreover, the F0 slopes were steeper in a compatible as opposed to a conflicting context, suggesting that tones coarticulate more with their adjacent tones in conflicting contexts than in compatible contexts, regardless of the prosodic position and information status of the tone-carrying syllable. In conclusion, the authors suggest that information status and prosodic structure both affect the strength and autonomy of tonal implementation, but do so in separate and independent ways.

Dick Smakman and Thomas de France, The acoustics of English vowels in the speech of Dutch learners before and after pronunciation training (288-301).The effect of phonetic training for L2 learners has been studied in various manners and languages. This paper focuses on the effects of an English phonetics course for university level native speakers of Dutch. Thirty-five female students participated in this experiment, in three groups, according to their initial pronunciation levels. They took pre-training and post-training tests, and the results were statistically analyzed. The authors of this paper found that the initially “best” group did not advance, but rather regressed in its production of the six tested English vowels (which are notorious for marking Dutch speakers). On the other hand, some of the participants progressed by 50%, above the average for such courses, while much variation appeared in other participants' results (improvement or regression). The authors suggest that more attention should be paid to the individualization of both phonetic training and course structure due to participants' diverse personal features.

Chaoju Tang, The use of Chinese dialects: Increasing or decreasing? Survey on the use of Chongqing dialect (302-210). China is known for its large number of spoken languages and dialects. However, the “common language” Putonghua, has become the main language used in it, causing diminished use of other languages and dialects, in particular among the younger generation. This paper focuses on the Chongqing dialect and its rate of use compared to some other Chinese dialects. Participants included students and non-students. The results varied by age and showed that Putonghua was dominant at workplaces, whereas the Chongqing dialect was used more at home. The author deduces that the dialect will apparently continue to be used at home for some time yet, and is not going to die out soon, though its rate of use is gradually changing. The author comments, however, that this finding, which differs from other surveys, may be due to the different (larger) number of participants in those surveys.

Alice Turk, Durational effects of phrasal stress (311-322). Phrasal stress (i.e., which syllable or word in a phrase is stressed), is studied from many angles. The paper begins with a definition of phrasal stress and summarizes the literature from several languages about speech stretch durations that affect phrasal stress (multiple syllables and constituent structure). Turk discusses various ways to describe the affected syllables (i.e., long syllables due to various structural conditions). The literature

provides contradictory evidence, which counters a suggested hypothesis (concerning the appropriate domain role, and different syllables magnitude). Turk considers the Multiple (Optional) Site hypothesis (which allows different durations in different conditions) as more appropriate for understanding the situation than the Continuous Domain hypothesis, though inter-language differences may relate to language-types (e.g., languages with short and long phonemic contrasts). Altogether, phrasal stress is quite complex, and involves many sites and mechanisms, some of which are optional.

Harry van der Hulst, The laryngeal class in RcvP and voice phenomena in Dutch (323-349). This is a theoretical phonology paper, although the theoretical considerations reflect physiological articulatory and phonetic processes. The author compares his theory of RcvP (Radical CV Phonology, van der Hulst 2005, and in preparation) with several other theories. His analysis leads to distinguishing vowels from consonants (as basic categories) and their features (elements), in the search for basic factors that produce contrasts within categories. The comparison refers, among other examples, to voicing/ devoicing in obstruent consonants in Dutch, vs. English, vs. French vs. Polish. Dutch /b, p/ are (theoretically) similar to English (both being Germanic languages), but they differ from English, probably due to the effect of (Romance) French. Another discussed issue is Dutch Final Obstruent Devoicing (FOD), examined here in relation to Government Phonology, and Dutch voicing assimilation. This theory involves the main consonantal category [fortis] vs. ø, which is enhanced in different ways in different languages. The paper shows the interaction between phonetic processes and well-defined phonemic representations. In Dutch, voicing assimilation and FOD are, thus, not just phonetic processes.

Jeroen van de Weijer, Affricates in English as a natural class (350-358). This last paper in the book is also rather theoretical. In the literature, affricates are described as stops, fricatives or complex segments. van de Weijer argues mainly against the "stop approach" (i.e., that affricates are stops). He finds that English affricates are neither stops nor fricatives by nature and behavior, and therefore form a different, natural segment class. He bases his claim on the analysis of occurrence restrictions in initial and final positions in English words. No consonant can precede affricates in initial position and only some (liquid and nasal) consonants can precede affricates in final position. This array differs from two other examined languages (Pengo in India, and Cimbrian German in Italy). Since "Phonotactic restrictions are usually regarded as the most secure kind of evidence for the phonological status of particular segment groups", the author concludes that according to this examination, affricates do form a separate consonant group – at least in English.

Altogether, this volume is rich in material in many areas of phonetics and phonology, and can benefit many readers interested in the discussed areas and languages.

References
Bailey, T. M. (1995). *Nonmetrical Constraints on Stress*. Unpublished PhD dissertation, University of Minnesota.
Bolinger, D. L. (1964). Around the edge of language: Intonation. *Harvard Educational Review*, 34, 282-296.
Heinz, J. (2007). *The Inductive Learning of Phonotactic Patterns*. Unpublished PhD dissertation, University of California, Los Angeles.
van Heuven, V. J., & Kirsner, R. S. (1999). Interaction of grammatical form and intonation: Two experiments on Dutch imperatives. In: Kager R., & van Bezoojen, R. (Eds.). *Linguistics in the Netherlands*, Amsterdam: John Benjamins, 165-183.
van Heuven, V. J., Neijt, A. H., & Hijzelendoorn, M. (1994). Automatische indeling van Nederlandse woorden op basis van etymologische filters [Automatic classification of Dutch words on the basis of etymological filters] *Spektator*, 23(4), 279-291.
van der Hulst, H. (2005). The molecular structure of phonological segments. In: Carr, P., Durand, J., & Ewen, C. (Eds.). *Headhood, Elements, Specification and Contrastivity*, Amsterdam: Jonhn Benjamins, 193-234.
van der Hulst, H. (in preparation). *Principles of Radical CV Phonology*, Ms. University of Connecticut
Kirsner, R.S. (1989.) Does sign oriented linguistics have a future? On sign, text and the falsifiability of theoretical constructs. In Y. Tobin (ed.) *From Sign to Text: A Semiotic View of Communication*. Amsterdam: John Benjamins. 161-178.
Kirsner, R. S., & van Heuven, V. (1988). The significance of demonstrative position in modern Dutch. *Lingua*, 76, 209-248.
Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*, Chichester: Wiley-Blackwell.
Laver, J. (1994). *Principles of Phonetics*, Cambridge: Cambridge University Press.
Lehiste, I. (1970). Suprasegmentals. Cambridge, MA: MIT Press.
Neef, M. (2005). Graphematics as part of a modular theory of phonographic writing systems, *Writing Systems Research*, 4, 214-228.

Edwin D. Lawson, Zinaida S. Zavyalova & Richard F. Sheil (2014): *Tatar First Names from West Siberia: An English and Russian Dictionary.*

San Diego, CA, Fredonia SUNY: HTComGroup,
Price: $14.99, €10.91, ISBN 9781-495373220; paperback.

Reviewed by: **Shlomit Landman[1], and Judith Rosenhouse[2]**
[1]Achva Academic College, Israel
[2]SWANTECH Ltd. and Technion I.I.T.
e-mail: shlomitlnd@gmail.com, judith@swantech.co.il

Edwin D. Lawson is a Professor Emeritus of Psychology, State University of New York, Fredonia. He was President (1995-1997) of the American Name Society and published over 150 books and articles, which included articles on Russian, Latvian, Lithuanian, and Azeri and Tatar names. Zinaida S. Zavyalova is an Assistant Professor in the Department of Cultural Studies and Social Communication, Tomsk Polytechnic

University. Her articles deal with a wide variety of topics in philosophy, linguistics, communication and onomastics. Richard F. Sheil was a Professor Emeritus of Music. He taught voice and choral conducting at the State University of New York for 30 years. He collaborated with Edwin D. Lawson and Farid Alakbarli to publish papers on the pronunciation and meaning of Azeri names and of the naming patterns of the mountain Jews of Azerbaijan. Also in collaboration with Edwin D. Lawson, he developed a website demonstrating the pronunciation of Russian, Estonian, Azeri and Tatar names. The book is dedicated to the memory of Professor Sheil (1919-2013). Several other local researchers (e.g., historians and linguists) assisted these authors in producing the dictionary.

This book is reviewed here because the subject of proper names is interesting phonetically as well as historically, lexically, etc. A few examples (below) show cross-language processes, expressed in phonetic features. The CD attached to the volume provides the orally recorded names in the dictionary which may be analyzed acoustically by researchers of Tatar language elements

This book is written in English, followed by a translation of each part into Russian. The Introduction (pp. xiii-xxx, English and Russian) contains a short explanation regarding the connections of politics and religion with onomastics. This connection is shown in the Russian influence on the onomastics of the Tatars of West Siberia during the last century. The Russian influence was mainly political, since the Tatars, a Muslim people, are known for their historical use of past names from Arabic, Persian, Turkish, Iranian, and Tatar.

The authors gathered information from three generations of 50 families from the city of Tomsk and, 50 families from villages in the Tomsk area, for a total of 799 persons. The Analysis identified evidence from nine time periods, Czarist (until 1917), Unsettled (1918-1920), Soviet (1921-1940), World War II (1941-1945), Post-War (1946-1953), Post-Stalin (1954-1964), Brezhnev (1965-1984), Gorbachev (1985-1990), and Post-Communist (1991-present). The results demonstrated the influence of Russian names and its naming system on the West Siberian Tatars. Most Tatars adopted Russian first names and patronymic names, some with the same initials as the original Tatar names and, some as a personal choice.

Three tables show the transcription from Russian to English, and from both languages to International Phonetic Alphabet (IPA) (pp. xxxi-xxxiv). The first table is the 'Russian to English pronunciation guide' in Cyrillic and Latin letters. The second one is the 'English to IPA pronunciation guide' and, 'Russian to IPA pronunciation guide' is the third table in Cyrillic and Latin (IPA) letters.

The dictionary (pp.1-63) includes 480 written entries, and comes with a CD featuring a native speaker's pronunciation of the names. Each entry contains an English and Russian spelling, gender and frequency in the sample, BBC-New York Times style of pronunciation, IPA style of pronunciation, language(s) of origin if not Tatar, meaning, and historical notes, if any.

Although the dictionary focuses on Tatar and Russian names, this book contains a large number of names that originate from Arabic, usually via Turkish. However, little

information is provided on these names in the book. As with other names, the transfer between cultures and languages involves a phonetic adaptation to the dominant language, which in this case is the Tatar or Russian language. Here are a few examples. The very first name (p. 1) *ab-'dul-ga-ni* uses the voiced stop /g/ instead of the voiced uvular fricative /ɣ/ of Arabic and the initial voiced pharyngeal /ʕ/ of Arabic is lost; *a-'di-ja* (p. 2) has lost the first consonant /h/ (in Arabic: /*hadijja*/); the Arabic name *ba-ga-'ʒat* (p. 6), “joy, delight,” is pronounced in Arabic with /h/, i.e., /*baḥʒat*/. However, in Russian foreign /h/ is usually pronounced as /g/, as we see in this name. In addition, a vowel is added after the /g/, unlike the cluster in Arabic. The name *sabir* (p. 44) “patient” in Arabic was originally produced with a pharyngealized /s/, i.e., /sˁ/; *vai-'du-la* (p. 52) is 'white + God'. (In Arabic, 'white' is associated with purity, freshness, and honor. Thus, this name means “God's honor”.) Nevertheless, in Arabic, this word is transcribed /bajadˁ-ulla/, i.e., two phonetic adaptations in a single word (name). It would be interesting to compare such processes in additional separate studies. This volume is therefore a recommended publication for researchers and scholars interested in onomastics, and will hopefully open up the way for further phonetic studies in this field.

CALL FOR PAPERS

The *Phonetician* will publish peer-reviewed papers and short articles in all areas of speech science including articulatory and acoustic phonetics, speech production and perception, speech synthesis, speech technology, applied phonetics, psycholinguistics, sociophonetics, history of phonetics, etc. Contributions should primarily focus on experimental work but theoretical and methodological papers will also be considered. Papers should be original works that have not been published and are not being considered for publication elsewhere.

Authors should follow the *Journal of Phonetics* guidelines for the preparation of their manuscripts. Manuscripts will be reviewed anonymously by two experts in phonetics. The title page should include the authors' names and affiliations, address, e-mail, telephone, and fax numbers. Manuscripts should include an abstract of no more than 150 words and up to four keywords. The final version of the manuscript should be sent both in .doc and in .pdf files to the Editor. It is the authors' responsibility to obtain written permission to reproduce copyright material.

INSTRUCTIONS FOR BOOK REVIEWERS

Reviews in the *Phonetician* are dedicated to books related to phonetics and phonology. Usually the editor contacts prospective reviewers. Readers who wish to review a book should address the editor about it.

A review should begin with the author's surname and name, publication date, the book title and subtitle, publication place, publishers, ISBN numbers, price, page numbers, and other relevant information such as number of indexes, tables, or figures. The reviewer's name, surname, and address should follow "Reviewed by" in a new line.

The review should be factual and descriptive rather than interpretive, unless reviewers can relate a theory or other information to the book which could benefit our readers. Review length usually ranges between 700 and 2500 words. All reviews should be sent in electronic form to Prof. Judith Rosenhouse (e-mail: judith@swantech.co.il).

ISPhS MEMBERSHIP APPLICATION FORM

Please mail the completed form to:

Treasurer:
Prof. Dr. Ruth Huntley Bahr, Ph.D.
Treasurer's Office:
Dept. of Communication Sciences and Disorders
4202 E. Fowler Ave. PCD 1017
University of South Florida
Tampa, FL 33620 USA

I wish to become a member of the International Society of Phonetic Sciences

Title:____ Last Name: _____ First Name: _____
Company/Institution:
_____
Full mailing address:
_____
_____
Phone: _____ Fax: _____
E-mail:
_____
Education degrees: _____
Area(s) of interest:
_____

The Membership Fee Schedule (check one):

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.000 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.000 per year
5. Libraries (plus overseas airmail postage)	$ 32.000 per year
6. Sustaining Members	$ 75.000 per year
7. Sponsors	$ 150.000 per year
8. Patrons	$ 300.000 per year
9. Institutional/Instructional Members	$ 750.000 per year

Go online at www.isphs.org and pay your dues via PayPal using your credit card.
☐ I have enclosed a cheque (in US $ only), made payable to ISPhS.

Date _____ Full Signature _____

Students should provide a copy of their student card

NEWS ON DUES

Your dues should be paid as soon as it convenient for you to do so. Please send them directly to the Treasurer:

Prof. Ruth Huntley Bahr, Ph.D.
Dept. of Communication Sciences & Disorders
4202 E. Fowler Ave., PCD 1017
University of South Florida
Tampa, FL 33620-8200 USA
Tel.: +1.813.974.3182, Fax: +1.813.974.0822
e-mail: rbahr@ usf.edu

VISA and MASTERCARD: You now have the option to pay your ISPhS membership dues by VISA or MASTERCARD using PayPal. Please visit our website, www.isphs.org, and click on the Membership tab and look under Dues for "paid online via PayPal." Click on this phrase and you will be directed to PayPal.

The Fee Schedule:

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.00 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.00 per year
5. Libraries (plus overseas airmail postage)	$ 32.00 per year
6. Sustaining Members	$ 75.00 per year
7. Sponsors	$ 150.00 per year
8. Patrons	$ 300.00 per year
9. Institutional/Instructional Members	$ 750.00 per year

Special members (categories 6–9) will receive certificates; Patrons and Institutional members will receive plaques, and Affiliate members will be permitted to appoint/elect members to the Council of Representatives (two each national groups; one each for other organizations).

Libraries: Please encourage your library to subscribe to *The Phonetician*. Library subscriptions are quite modest – and they aid us in funding our mailings to phoneticians in Third World Countries.

Life members: Based on the request of several members, the Board of Directors has approved the following rates for **Life Membership** in ISPhS:

Age 60 or older:	$ 150.00
Age 50–60:	$ 250.00
Younger than 50 years:	$ 450.00