Linguistic & Paralinguistic Phonetic Variation
in Speaker Recognition & Text-to-Speech Synthesis

Susanne Schötz (susanne.schotz@ling.lu.se)
Department of Linguistics and Phonetics, Lund University

ABSTRACT

Phonetic variation, and especially prosodic variation, which is often paralinguistic in nature has gradually attracted more attention among speech researchers and speech scientists as one of the possible solutions to problems with automatic speaker recognition (ASrR) and text-to-speech synthesis (TTS) systems. This paper presents a brief overview of approaches to phonetic variation in ASrR and TTS, beginning with attempts to classify linguistic and paralinguistic phenomena in speech. Also, some of the problems related to paralinguistic phonetic variation and attempted solutions are discussed.

1 Introduction

One of the major obstacles to overcome when trying to improve existing speaker recognition and text-to-speech systems is related to prosody. The prosody models of today are still far from perfect, and as paralinguistic information in speech is mainly signalled with prosodic cues, the systems of today are also unable to effectively and reliably recognize and generate speaker specific qualities like age, sex, emotions and attitudes. Another problem for speech researchers is related to the confusing terminology associated with prosody. Linguists have made a distinction between linguistic and paralinguistic, while phoneticians traditionally have preferred to draw the line between segments (vowels and consonants) and prosody.

In this paper some of the approaches to the problems associated with linguistic, prosodic and paralinguistic phonetic variation are presented and discussed. Of the following sections, 3.1 and 4.1 are based mainly on Furui (1996, 1997) and Gish & Schmidt (1994), while sections 3.2 and 4.2 are based on Dutoit (1997), Klatt (1987) and Carlson & Granström (1997).

2 Classifications of phonetic variation

There are a number of ways to classify phonetic variation in speech, and it is not always agreed upon what the different categories are and which aspects they should include. Obviously, different typologies are created for different purposes, but categories often overlap and several features may belong to more than one category. This section presents some of the different classes that have been used for describing phonetic and paralinguistic variations, and also provides examples of the various categories.

2.1 Linguistic or paralinguistic

A distinction typically made by linguists and many speech researchers is one that divides speech into linguistic information, i.e. the arbitrary language code used intentionally by the speaker for communication on one hand, and all other information on the other. Speech signals necessarily contain other information besides linguistic. Such information varies as a function of the speaker, the listener/s and the communicative situation, and is referred to as *paralinguistic*, *extra-linguistic* or *non-linguistic* in the literature. In Saussure's terminology paralinguistic phenomena would rather be 'parole' than 'langue' (Traunmüller 2001).

Roach et al (1998) define *paralinguistic* features as those used intentionally by the speaker, and *non-linguistic* features as those that cannot be used intentionally, such as age, sex, state of health etc. *Non-linguistic* features are further classified into *individual variation*, due to the physiology (size, weight) and histology (age) of the vocal tract, which affect the phonation and resonance of the speech, and *reflexes*, that are involuntary reactions to an emotional state and include clearing the throat, sniffs, yawns, laughs, cries and audible breathing.

Carlson (2002) uses the term *extralinguistic* for inhalation, exhalation, smacks and hesitation sounds, and Carlson & Granström (1997) refer to attitudes and emotions as *extralinguistic*.

Traunmüller (2000, 2001) suggests that information in speech can be categorised into *linguistic* (message, dialect, sociolect, speaking style), *organic* (age, sex, pathology), *expressive* (adaptation to environment, emotion and attitude) and *perspectival* (distance, direction, transmission channel etc.).

In Lindblad (1992) the term *paralinguistic* is used to denote speech sounds signalling other information than linguistic, such as emotions, attitudes, age, sex, dialect and sociolect.

Laver (1980) uses the term *paralinguistic* for signals of affective information through tone of voice and conversational interaction regulations, and the term *extralinguistic* for voice qualities identifying the individual speaker.

Marasek (1997) refers to Laver (1991, 1994) when describing speech as a multi-layer medium; the *linguistic* layer for semantic information and phonetic representation, the *paralinguistic* layer for *non-linguistic* and non-verbal information about the speaker's attitudes, emotions, regional dialect and sociolect, and the *extralinguistic* layer for physical and physiological (including organic) features, such as the speaker's sex, age and habitual factors.

According to Quast (2001) information in the speech signal is communicated on three different channels; the *linguistic*, *paralinguistic*, and *extralinguistic* channels. The verbal content, the actual meaning of the words, is thought of as linguistic information. Information about the speaker's basic state, including the size of the speaker's body and vocal tract, and the culture of the speaker, such as the use and range of pitch movements throughout an utterance. The *paralinguistic* channel carries information about momentary changes in the usual (*extralinguistic*) baseline, such as whispering in a situation that calls for silence, or expression of emotions.

Mixdorff (2002) subdivides *prosodic* information in speech into three categories. *Linguistic* information includes lexical stress, sentence modality (question vs. non-question), focus structure and segmentation, while *paralinguistic* information comprises speaker attitude, intention, dialect and sociolect, and *non-linguistic* deals with emotions and health.

As can be seen above as well as in table 1 below, the terminology is rather confusing.

Table 1. Some classifications of paralinguistic information in speech found in the literature.

	paralinguistic	**extralinguistic**	**non-linguistic**	**(intra)linguistic**
Carlson 2002		inhalation, exhalation, smacks, hestitation sounds		
Carlson & Granström 1997		attitudes, emotions		
Laver 1980	affective information	voice qualities identifying the speaker		
Linblad 1992	emotions, attitudes, age, sex, dialect, sociolect, (health?)			
Marasek 1997	non-linguistic and non-verbal information; attitude, emotions, dialect, sociolect	physical & physiological features; age, sex, habitual factors		
Mixdorff 2002	speaker attitude, intention, dialect, sociolect		emotions and health	
Quast 2001	momentary changes; whispering, emotions	the speaker's basic state; physical, physiological (body & larynx size)		
Roach 1998	intentional; voice qualities (modal, falsetto, breathy voice etc.) and voice qualifications (non-linguistic vocal effects (laughing, sobbing, tremor etc.)		unintentional; age, sex, health	
Traunmüller 2000, 2001	organic; age, sex, health, and expressive; emotion, attitude, adaptation to environment	perspectival; distance, direction, transmission channel		linguistic: dialect, sociolect, speaking style

In this paper, unless explicitly stated otherwise, the term linguistic is used when referring to the arbitrary code of language, while the term paralinguistic is used to denote all other aspects of speech, such as speaker age, sex, health condition, emotional state and attitude.

2.2 Paralinguistic or prosodic

How and where to draw the line between *paralinguistic* and *prosodic* features is also disagreed upon.

Mozziconacci (2002) describes *prosodic* cues as fulfilling a linguistic function (discourse and dialogue structuring, focus signalling) but also providing information about the speaker's sex, age, physical condition as well as the speaker's view, emotion and attitude toward the topic, dialogue partner or situation.

Roach et al (1998) refer to the work of Crystal (1969) when making the following distinction: *prosodic* features are characterized by variations in pitch, loudness, duration and silence, and *paralinguistic* features are vocal but independent of pitch, loudness and duration for their identification. The categories proposed by Crystal & Quirk (1964) and Laver (1980) are presented in Roach et al (1998) as gradient, with *prosodic* features unambiguously signalling linguistic information at one end, and features such as voice quality and clicking noises on the other *paralinguistic* end of the scale. A further division of *paralinguistic* features is made into *voice qualities*, due to different modes of phonation, such as modal voice, falsetto, whisper, creak, harshness and breathiness, and *voice qualifications*, which are non-linguistic vocal effects such as laughing, giggling, tremulousness, sobbing and crying. *Prosodic* features are further divided into tempo, prominence, pitch range, rhythmicality, tension, pause and intonation. (Roach et al 1998)

Quast (2001) argues that speech incorporates a verbal and a nonverbal communication channel, where the verbal part is represented by words and *the nonverbal channel is carried by the prosody*, i.e. the stress and intonation patterns of the utterance, *and* holding information about the speaker's physical state, emotions, the attitude towards the object of the conversation etc.

Since prosody can be used to signal both linguistic and paralinguistic information, and as prosody and paralinguistic phonetics share most of their phonetic correlates, one may regard paralinguistic phonetics as a subset to prosody. This can be seen in figure 1 below.

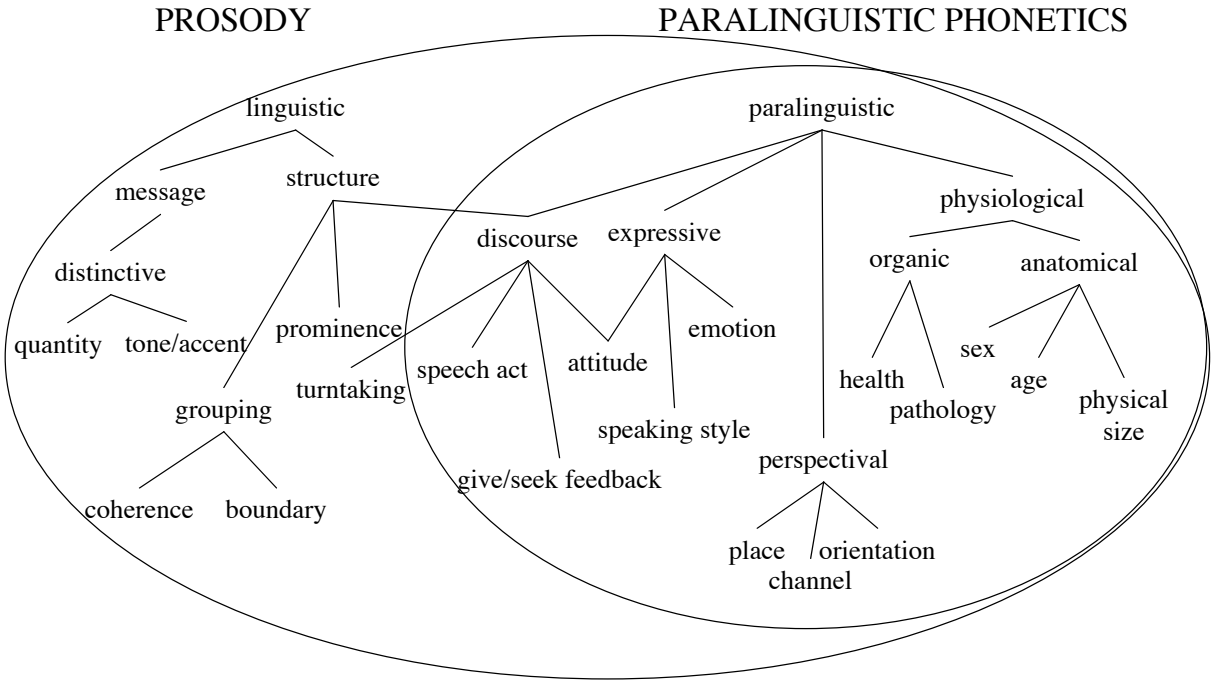

Figure 1. Some functions of prosody and paralinguistic phonetics.

2.3 Other categories

In addition to the categorisations mentioned above, there are a number of other perspectives and approaches to studies of phonetic variation. Phonetic studies are usually carried out in the *articulatory* (i.e. production), *acoustic* or *perceptual* dimensions, but also in the dimensions *time*, *frequency* (spectral) and *intensity* (amplitude). Speech can be further divided into variations regarding *intra-speaker - inter-speaker*, *segmental* (phonemic, microprosodic) - *suprasegmental* (prosodic), *long-term - short-term*, and *language-specific* (or cultural) - *universal*. Moreover, a number of speech variation categories overlap, and many aspects can be placed in more than one category, e.g. speaking styles, which can be either linguistic (formal, casual) or paralinguistic (baby talk, foreigneese).

3 Linguistic phonetic variation

When we speak, we use the lexicon and grammar of a language, and we use one of several dialects that a language normally consists of. In addition to this, foreign accents can be noticed in the phoneme inventory and in the prosodic patterns of non-native speakers. Moreover, every speaker's idiolect differs more or less from other speakers of the same dialect. Apart from between-speaker variation, there is also within-speaker linguistic phonetic variation, e.g. the same speaker may use different speaking styles, like formal, clear, casual or sloppy speech, depending on the listener and the situation. In clear speech words are pronounced in canonical form. However, when speaking in a more casual style, words or phrases are often simplified, leading to coarticulation, assimilation or reduction of many speech sounds. These simplifications are often more dramatic than we are aware of, since humans seldom have any problems understanding reduced forms. E.g. the Swedish three syllable phrase 'det är ett' [de:t æ:r et] (it is a) is often reduced to a single syllable [de:t] in casual speech. Linguistic phonetic variation is language specific and like language itself, it is arbitrary. Every language has its own rules for phonetic variation. (Traunmüller 1997)

3.1 Linguistic phonetic variation in ASrR

Unlike automatic speech recognition, ASrR does not have to handle linguistic phonetic variation very often. Speakers seldom change their dialects, unless when trying to fool the ASrR system by disguising their own voices or imitating other voices. Although methods in ASrR can be either text-dependent or text-independent, no methods to explicitly resolve linguistic phonetic variation are required even for text-dependent systems, since the key words or sentences required are the same text for both training and recognition. If a speaker moves away from his/her native area (e.g. from England to South Africa, or from Malmö to Stockholm), some of the speaker's speech habits may adapt to the new dialect leading to changes in the speaker's allophone inventory and/or prosodic patterns. This is usually a gradual process over a long period (a few years) of time. Normalisation and adaptation techniques can be used to handle such variation (see section 4.1 below).

3.2 Linguistic phonetic variation in TTS

Phonetic variation in a typical TTS system is handled by the natural language processing (NLP) module, where the phonetic transcription (including intonation and rhythm) of the input text is produced, and in the digital signal processing (DSP) module, where the information received from the NLP module is transformed into actual speech. The NLP module, which typically comprises a text analyser, a letter-to-sound (LTS) module and a prosody generator, operates on several linguistic levels, including morphology and syntax.

Since linguistic phonetic variations are arbitrary and language-specific, developers adapt their TTS systems to the languages (and dialects) they want to produce. Typical segmental variations in fluent speech include consonant cluster simplification, assimilation, heterophonic homographs, and deletion or reduction in duration and spectral space of unstressed segments. Strategies for handling segmental variation include dictionary-based methods, which store a maximum of phonological knowledge into a

lexicon, or rule-based methods, where most of the phonological competence is transferred into LTS rules.

Perhaps even more important than segmental variations are prosodic variations, since they are processed on all linguistic levels: in the lexicon, syntax, semantics and pragmatics. Variations in intensity, fundamental frequency (F_0) and duration are predicted and generated in the prosody generator. Here one of various prosody models[1] is implemented along with rule-based, corpus-based and/or statistical methods, including HMMs and classification and regression trees (CARTs).

To increase the naturalness and intelligibility, phonetic postprocessing (e.g. spectral and prosodic smoothing at splice points) in the DSP module is often necessary. In concatenation synthesis systems PSOLA (pitch-synchronous overlap-add) methods have outperformed linear predictive coding (LPC) methods in manipulating prosodic parameters, but a genuine natural prosody generator has yet to be developed. Given the complexity of this task, some researchers have settled for "acceptable neutral" prosodic variations in the speech produced by their TTS systems. Other approaches include corpus-based methods with non-uniform unit selection, where the natural prosody is preserved whenever possible and model-based prosody is imposed otherwise (Möbius 2000), and limited target domains where the reduced vocabulary size and variety of sentence structures have increased both the segmental and prosodic naturalness (van Santen et al 2000).

4 Paralinguistic phonetic variation

In his introduction, Laver (1980) quotes Quintilian (c.III, Book XI), who wrote "The voice of a person is as easily distinguished by the ear as the face by the eye". Humans have no trouble identifying an individual speaker or recognizing paralinguistic information in speech. The question asked by many speech scientists is: How do we do it? Paralinguistic features are often difficult to measure and isolate acoustically from other features or information because they appear in more than one dimension. Moreover, paralinguistic distinctions are often gradient (e.g. age, emotions etc.) and there seem to be both inter- and intra-speaker differences. These are some of the difficulties automatic speech recognition and TTS systems have to face. Could it be that humans use individual combinations of acoustic parameters as cues for paralinguistic features? Lately, much research has been devoted to such questions. Results so far indicate that spectral and durational information may be of more importance than pitch and energy (Furui 1997, Schötz 2001).

In order to avoid confusion with paralinguistic feature classifications, this section uses only the terminology proposed by Traunmüller (2000), which divides all non-linguistic information into expressive, organic and perspectival. *Expressive* variation is realized as variation in speech rate, loudness, liveliness and voice quality, and is used to signal emotion, attitude and adaptation to the environment. This variation may be voluntary or involuntary, and often has drastic effect on the acoustics. *Organic* variation is caused by differences in the size and form of the speech organs, and is primarily related to speaker age and sex. It is most easily observed between speakers or within a single speaker, as he/she grows older, and variations include average pitch and spectral quality of speech sounds. *Perspectival* variations arise on the way from the speaker to the listener, and include variation in loudness, acoustic effects (e.g. echo), and external background or transmission channel noise.

4.1 Paralinguistic phonetic variation in ASrR

In ASrR speaker-specific information included in the speech wave is used to identify a speaker from a group of registered speakers (speaker identification) or verify the claimed identity of a speaker (speaker verification). Typical ASrR systems consist of two components: one for training, where data from the speakers to be identified are collected, and one for testing, where an input utterance is compared to the training data and the actual identification is made. Feature parameters often used in ASrR systems are F_0 for intonation, along with short-term and long-term spectra for implicitly

[1] For a more detailed review of prosodic models, see Dutoit (1997).

capturing organic vocal tract and voice quality characteristics. LPC derived cepstral coefficients and their regression coefficients are currently the most commonly used short-term spectral measurements.

Text-dependent ASrR systems use template-matching techniques with dynamic time warping (DTW) and hidden Markov model (HMM) based methods. Short-term spectral feature vectors are used to represent each utterance, and DTW algorithms are applied to align the test utterance with each reference template in time. Then the overall distance between the test utterance and the template is calculated. In HMM approaches the statistical variation in spectral features is modelled, and each utterance represented as a sequence of subword units.

Text-independent systems cannot match phonemes or words when trying to recognize a speaker, so they have to rely on other methods. Long-time average spectra (LTAS), weighted cepstral distance measures and multivariate auto-regression (MAR) models are often used to capture the spectral and dynamic characteristics of individual speakers. Other methods include vector quantization (VQ), ergodic HMMs and speech-recognition-based methods. VQ codebooks, where compressed data representing speaker-specific features are stored in a small number of vectors for each speaker, are used to vector-quantize an input utterance for recognition decision. Ergodic HMMs use the same structure as VQ-based methods, but the VQ codebook is replaced by an ergodic HMM with only a few broad phonetic category states.

Text-prompted speaker recognition with speaker-specific phoneme models, where the user is prompted to speak a randomly selected utterance, has been used to cope with problems with impostors using recorded voices and people who do not like to utter their password within hearing range of other people.

Expressive and *organic* variations in speech include age, short-term illness, emotional state, attitude, deliberate disguise or imitation etc. Production oriented methods can be used to synthesize speaker-specific features in terms of voice quality, vocal tract length, speaking habits etc., but these approaches are still considered rather unreliable due to problems with production parameter extraction and lacking phonetic knowledge (Blomberg & Elenius 2002).

Some *perspectival* variations in the speech signal characteristics arising from differences in recording and transmission conditions, from background noise or crosstalk, can be reduced by instructions to the speaker on where to stand (e.g. 20 cm from the microphone), when and how to speak (e.g. in a normal tone) etc. Others are handled with normalisation techniques, adapting the verification threshold and reference model for each speaker to such variation. Two normalisation approaches have been tried for this purpose; parameter-domain spectral equalisation methods, where cepstral coefficients are averaged over the duration of an entire utterance and the averaged values are subtracted from the cepstral coefficients to compensate for additive variation in the log spectral domain, and distance (similarity or likelihood) methods. However, both approaches have encountered severe problems. HMM methods have recently been tried for noisy speech, where a clean speech HMM is combined with one or several noise HMMs.

4.2 Paralinguistic phonetic variation in TTS

Paralinguistic variation is crucial to human sounding synthetic speech, but has proven to be one of the most difficult tasks to solve in TTS systems. The last decades a number of approaches to obtaining more natural speech have emerged. Concatenation synthesis inherits the voice personality along with other long-term and short-term paralinguistic features from the speaker who recorded the speech database, but this does not necessarily lead to more normally sounding speech, since the recorded features often are difficult to change. In theory articulatory and formant TTS systems should be able to produce any paralinguistic variations, but lacking adequate paralinguistic models and methods to handle these features, they are still unable to produce natural sounding speech.

Because of these difficulties, *expressive* variations like emotions are generally held at a minimum in current TTS-systems. However, some advances have been made lately. Murray and Arnott (1993) found that the voice parameters affected by emotion were of the three types voice quality, utterance timing and utterance pitch contour. Cahn (1990) has experimented with an ”affect editor”, which uses an abstract model of emotional speech along with generation instructions to produce recognizable and sometimes even natural sounding emotions with formant synthesis. Boula de Mareüil et al (2002) have

synthesized six emotions (anger, disgust, fear, joy, surprise and sadness) in three languages (English, French and Spanish) using corpora and morphing techniques.

Organic variation is difficult to manipulate, especially in concatenation synthesis. Klatt (1987) reports of storing detailed formant data of a man's and a woman's voice for creating other male, female and child voices using formant synthesis, but points out that although it is possible to modify parameters like F_0, F_0 range, spectral tilt, glottal open quotient, and breathiness, truly feminine voice quality remains elusive in TTS systems. In voice conversion systems, where a source speaker's speech is modified to sound as if it was spoken by a target speaker, spectral transformations have been performed with PSOLA techniques (Gutiérrez-Arriola et al 2001) and methods with LPC-based algorithms and VQ have also been used (Kain & Macon 1998).

TTS systems need not to worry about *perspectival* variations, since such variation arises after the speech has left the loudspeaker/headphones of the system, but could perhaps utilize variations in vocal effort to compensate in a natural sounding way for very noisy conditions.

5 Problems with paralinguistic phonetic variation

One of the main reasons for the limited success with paralinguistic phonetic variation in ASrR and TTS systems is the absence of adequate models for prosodic and paralinguistic features. A number of prosody models are available for speech technology applications today, but state-of-the-art systems using these models still fail to produce natural-sounding prosody. Other possible reasons for problems with paralinguistic features include insufficient knowledge of the human hearing capability and insufficient research attempts to handle paralinguistic features. This section attempts to highlight some of the problems with paralinguistic information regarding speaker recognizing and speech synthesis.

5.1 Problems with paralinguistic phonetic variation in ASrR

Problems in ASrR systems are related to difficulties in automatic extraction and modelling of suprasegmental (prosodic and paralinguistic) features due to their high inter-speaker and intra-speaker variability. Perspectival features are also highly variable, leading to similar extraction and modelling problems.

Normalisation and adjustment techniques are faced with a number of problems. Parameter-domain normalisation methods unavoidably remove some important speaker-specific features, making them inappropriate for short utterances. Likelihood normalisation methods have to deal with problems concerning reference speakers or "cohort speakers" and opposite-gender impostors. Since the computational cost for calculating condition probabilities is high, the number and type of reference speakers has to be carefully chosen. If only same-gender speakers are chosen as reference speakers, the system becomes vulnerable to opposite-gender impostors. Also, some normalisation models are unable to differentiate between highly dissimilar speakers, as they neglect the absolute deviation between the claimed speaker's model and the input speech. (Furui 1997)

5.2 Problems with paralinguistic phonetic variation in TTS

The main problem with TTS systems today is that they do not sound acceptably natural in terms of segmental and suprasegmental phonetic variability. 1987 Klatt argued that acoustic phonetic analysis of phonemes is not yet sufficiently detailed for synthesis purposes, and this still holds true. Representations of phonemes, allophones, stress, and syntactic symbols are present in most current TTS systems, but more information about the phonetic and paralinguistic quality of speech is needed in order to produce naturally sounding synthetic speech, including qualities such as age, sex, speaking style etc. Naturalness should, however, not be confused with intelligibility (Klatt 1987). Although intelligibility of the best current TTS systems is very good, listeners immediately recognize that speech generated by TTS is not human. It is commonly believed that lack of natural prosody is one of the main reasons for this (van Santen 1997). Other problems concern the questions how to adequately formalize the relationship between paralinguistic and prosodic phonetic variation and syntax,

semantics and pragmatics and how to predict paralinguistic intonation, duration and spectral qualities from abstract patterns.

6 Conclusions, solutions and discussion

This section is not intended to provide the reader with viable answers to the problems concerning paralinguistic phonetic variation in TTS and ASrR systems, but rather proposes some possible solutions already suggested by speech scientists.

There is an acute need for an international multidisciplinary categorization system with consistent terminology describing paralinguistic and prosodic features. This, along with an exhaustive and unambiguous paralinguistic transcription system would undoubtedly increase the possibilities and accelerate the process to overcome many of the problems mentioned in the last section. Given an adequate typology and a corresponding transcription system, existing and new speech corpora could be paralinguistically annotated. Some emotional speech corpora have already been labelled manually (Roach et al 1998, Roach 2000, Gustafson-Capková 2001), and in the near future it should be possible to achieve high speed and accuracy in automatic methods for tagging as well as for retrieving all sorts of paralinguistic information (Carlson & Granström 1997). Besides being used directly in speech applications, automatically retrieved paralinguistic information could be studied further in order to construct theoretical and computational models for paralinguistic features comprising feature parameters as well as acoustic and perceptual correlates. A better understanding of the acoustic theory of speech production, would lead to better models of the larynx and the vocal tract, and, even more importantly, to a better understanding of human listeners' perceptual behaviour in terms of acoustic spectral and waveform features (Klatt 1987). Existing models for prosody may offer guidance when constructing paralinguistic models, since several features are shared.

Techniques for the phonetic description of prosody (Dutoit 1997) and paralinguistic features including age, sex, vocal effort (Traunmüller 1996, 1997) and emotions (Roach 2000) are already available, but there is definitely more work to be done in this area.

In the mean time, restricted target TTS domains (van Santen et al 2000), such as children's books, weather reports, banking services etc., with reduced vocabulary, sentence structures and speaking styles may serve as intermediate solutions for handling paralinguistic variation.

Evidently, more research is needed in order to achieve the goal of controlling paralinguistic features in speech technology applications. Progress has to be made in linguistics, phonetics and engineering, but a more multidisciplinary approach is advertised.

Production oriented systems is one example where different research fields have joined forces. Physical and articulatory knowledge is merged with speech technology methods in order to extract speaker-specific information. Features are separated and then described and trained on separately. This approach implies the combining of automatic training with knowledge from speech analysis and synthesis, and possible development of joint systems for ASrR and TTS, based on the same production model. (Blomberg & Elenius 2002)

Synthesized speech has been used for some time in phonetic laboratories, and phonetic models have been implemented in speech technology applications. The gap between these two disciplines is closing, enabling phoneticians and speech technologists to join forces in the quest for solutions to problems in speech science.

Finally, our attitude towards the distinction between human and computer-generated speech should perhaps be taken into account. Do we really want computers to speak just like us? Does computer speech not signal paralinguistic features of its own (e.g. this is a computer speaking)? Maybe we just want synthetic speech to sound acceptably natural, not perfectly natural. In that case some of the unnatural sounding and "robot-like" paralinguistic features should be preserved. In time, we may even come to think of them as charming.

7 References

Blomberg, Mats & Elenius, Kjell. 2002. *Automatisk igenkänning av tal*. Institutionen för tal, musik och hörsel, KTH.

Boula de Mareüil, P, Célérier, P & Toen, J. 2002. *Generation of Emotions by a Morphing Technique in English, French and Spanish*. In Bel, Bernard & Marlien, Isabelle (Eds.) Speech Prosody 2002. Proceedings, Aix-en-Provence, France.

Carlson, Rolf & Granström, Björn. 1997. *Speech Synthesis*. In Hardcastle & Laver (Eds.). The Handbook of Phonetic Sciences. Blackwell Publishers Ltd, Oxford (pp 768-788).

Carlson, Rolf. 2002. *Dialogsystem*. Slide presentation, Speech Technology, GSLT, Göteborg, October 23 2002. http://www.speech.kth.se/~rolf/gslt/GSLT021023_dialog.pdf.

Dutoit, Thierry. 1997. *An Introduction to Text-to-Speech Synthesis*. Dordrecht. Kluwer Academic Publishers.

Dutoit, Thierry. 1997. *High-quality text-to-speech synthesis: an overview*. In Journal of Electrical & Electronics Engineering, Australia. Special Issue on Speech Recognition and Synthesis, vol. 17 n°1 (pp. 25-37).

Furui, Sadoki. 1996. *Speaker Recognition*. In Cole, Ronald A. et al (Eds). Survey of the State of the Art in Human Language Technology (chapter 1.7). http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Furui, Sadoki. 1997. *Recent Advances in Speaker Recognition*. In Proceedings of AVBPA 1997 (pp 237-252).

Gustafson-Capková, Sofia. 2001. *Emotions in speech: Tagset and Acoustic Correlates*. Term paper in Speech Technology 1, Swedish National Graduate School of Language Technology (GSLT). Stockholm University. Department of Linguistics.

Gutiérrez-Arriola, J. M. et al. 2001. *A new Multi-Speaker Formant Synthesizer that applies Voice Conversion Techniques*. In Proceedings of Eurospeech 2001. Aalborg, Denmark.

Kain, A & Macon, M. W. 1998. *Spectral Voice Conversion for Text-to-Speech Synthesis*. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Vol. 1 1998 (pp. 285-288).

Laver, John. 1980. *The phonetic description of voice quality*. Cambridge University Press.

Lindblad, Per. 1992. *Rösten*. Lund. Studentlitteratur.

Marasek, Krzysztof. 1997. *EEG & voice quality* (Tutorial). http://www.ims.uni-stuttgart.de/phonetik/EGG/frmst1.htm.

Möbius, Bernd. 2000. *Corpus-based speech synthesis: methods and challenges*. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart). AIMS 6 (4) (pp 87-116).

Mixdorff, Hansjörg. 2002. *Speech Technology, ToBI, and Making Sense of Prosody*. In Bel, Bernard & Marlien, Isabelle (Eds.) Speech Prosody 2002. Proceedings, Aix-en-Provence, France.

Mozzioconacci, S. J. 2002. *Prosody and Emotions*. In Bel, Bernard & Marlien, Isabelle (Eds.) Speech Prosody 2002. Proceedings, Aix-en-Provence, France.

Quast, H. 2001. Automatic Recognition of Nonverbal Speech: An Approach to Model the Perception of Para- and Extralinguistic Vocal Communication with Neural Networks. Thesis. University of Göttingen.

Murray, Ian R. & Arnott, John L. 1993. *Toward the simulation of emotions in synthetic speech: A review of the literature on human vocal emotion.* Journal of Acoustical Society of America 93 (pp 1097-1108).

Roach, P, Stibbard, R, Osborne J, Arnfield S, Setter, J. 1998. *Transcription of Prosodic and Paralinguistic Features of Emotional Speech.* Journal of the International Phonetic Association (1998) 28 (pp 83-94).

Roach, Peter. 2000. *Techniques for the Phonetic Description of Emotional Speech.* In Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland. September 2000 (pp 53-59).

Roach, Peter. 2000. *The Emotion in Speech Project.* http://www.rdg.ac.uk/AcaDepts/ll/speechlab/emotion/.

Schötz, Susanne. 2001. *Röstens ålder – en auditiv och akustisk studie*. M.A. thesis in Phonetics. Lund University. Department of Linguistics and Phonetics.

Traunmüller, Hartmut. 1996. *Manipulations in speaker age and sex.* http://www.ling.su.se/staff/hartmut/manipul.htm.

Traunmüller, Hartmut. 1997. *Paralinguistic Variation in Speech and How to Handle it in Speech Technology.* http://www.ling.su.se/staff/hartmut/paraling.htm.

Traunmüller, Hartmut. 1997. *En tur i fonetikens marker*. http://www.ling.su.se/staff/hartmut/tur.htm.

Traunmüller, Hartmut. 2000. *Evidence for demodulation in speech perception.* http://www.ling.su.se/staff/hartmut/demod.pdf (contribution to ICSLP 2000).

Traunmüller, Hartmut. 2001. *Paralinguale Phänomene*. To appear in Ammon, A. Dittmar, N & Mattheier, K. J. (Eds.). Soziolinguistik – Ein Internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft. 2nd ed. Berlin/New York.

van Santen, Jan et al. 2000. *When will synthetic speech sound human: Role of rules and data*. In Proceedings of ICSLP 2000, Beijing.

van Santen, Jan. 1997. *Prosodic modeling in text-to-speech synthesis*. In Proceedings Eurospeech 1997, Rhodos, Greece.