

RAPORT merytoryczny szczegółowy nr 2

opracowania z roku 2012

Autorzy raportu i wykonawcy opisanych prac: Katarzyna Klessa, Maciej Karpiński, Agnieszka Wagner, Magdalena Oleśkowicz-Popiel oraz (tylko w zakresie rozdz. 5: Piotr Francuzik, Mariusz Owsiany)

(dotyczy realizacji projektu rozwojowego nr O R00 0170 12 pn. Pozyskiwanie i przetwarzanie informacji słownych w militarnych systemach zapobiegania oraz zwalczania przestępczości i terroryzmu)

Spis treści

1 Cechy paralingwistyczne. Parametryzacja, możliwości analizy cech na podstawie dostępnych korpusów.....	3
1.1 Cechy parajęzykowe a identyfikacja i rozpoznawanie głosu/ mówcy.....	3
1.2 Propozycje zestawów tagów do anotacji cech paralingwistycznych.....	14
1.3 Właściwości akustyczne pauz wypełnionych w mowie spontanicznej.....	16
1.4 Analiza zawartości bazy Jurisdict pod kątem badania cech paralingwistycznych.....	17
1.5 Analiza porównawcza cech indywidualnych wypowiedzi w mowie natywnej i nienatywnej.....	21
2 Kontynuacja prac nad analizą emocji.....	23
2.1 Cel eksperymentu.....	23
2.2 Wstępne testy z wykorzystaniem Emocjonalnej bazy Danych Politechniki Łódzkiej.....	23
2.3 Analiza percepcyjna.....	25
2.4 Analiza akustyczna.....	25
2.5 Rozpoznawanie emocji - modelowanie statystyczne.....	27
3 Korpus nagrań „Paralingua”.....	27
3.1 Struktura korpusu.....	27
3.2 Dialogi zadaniowe - podzbiór „Dial”.....	27
3.3 Mowa nacechowana emocjonalnie - podzbiór „Emo”.....	34
3.4 Analiza zawartości bazy.....	35
4 Annotation System. Oprogramowanie do anotacji nagrań mowy spontanicznej. .	38
4.1 Informacje ogólne.....	38
4.2 Architektura programu.....	38
4.3 Instalacja i uruchomienie programu.....	40
4.4 Rozpoczęcie pracy.....	40
4.5 Operacje na segmentach i warstwach anotacji.....	44
4.6 Przykładowe specyfikacje anotacji w programie i przeprowadzone analizy. .	56

Fundacji Uniwersytetu im. A. Mickiewicza

1 Cechy paralingwistyczne. Parametryzacja, możliwości analizy cech na podstawie dostępnych korpusów

1.1 Cechy parajęzykowe a identyfikacja i rozpoznawanie głosu/ mówcy

Rozdział dotyczy cech, które można wyodrębnić z wypowiedzi mówionej, a które - na podstawie przeglądu publikacji uznano za przydatne w identyfikacji (i weryfikacji) mówcy („właściciela” głosu), jak i jego stanu. Obejmuje:

- ogólne wprowadzenie;
- listę („mierzalnych wprost”) parametrów akustycznych sygnału, w znacznej części możliwych do automatycznej ekstrakcji (niektóre metody będą prawdopodobnie wymagały np. wyodrębniania fragmentów sygnału przez eksperta); niektóre cechy wymieniane są w zestawieniach - wtedy, gdy istnieją szczególne metody, oparte na wykorzystaniu takich zestawień cech (nie tylko jako elementów szerszego wektora cech);
- listę parametrów wypowiedzi, które mogą być lub powinny być „mierzone” w odniesieniu do jej struktury językowej;
- spis często stosowanych algorytmów i metod statystycznych stosowanych w systemach rozpoznawania i weryfikacji głosu/mówcy oraz ustalających cechy mówcy na podstawie cech głosu;
- zestawienie informacji pochodzących z najnowszego dostępnego artykułu, w którym porównywano skuteczność metod rozpoznawania mówcy;
- konkluzję i propozycje wdrożeniowe i badawcze;
- spis literatury przedmiotu, obejmujący około 150 pozycji (zawiera on również prace, których nie cytowano wprost w niniejszym tekście).

(Uwaga: W większości wypadków pozostawiono angielskie nazwy parametrów, żeby uniknąć nieścisłości.)

Informacje dotyczące identyfikacji stanów emocjonalnych potraktowano marginalnie, gdyż dotyczy ich inna część raportu. Z drugiej strony, należy pamiętać, że w praktyce należy unikać oddzielnego analizowania tych grup zmiennych, gdyż każdy stan emocjonalny może w szczególny sposób wpływać na różne parametry wypowiedzi.

Niekiedy autorzy tekstów rozróżniają rozpoznawanie głosu i mówcy, lecz w większości wypadków metody są podobne. Z reguły rozróżnia się również identyfikację od rozpoznawania, chociaż wydaje się, że procesy te różnią się bardziej na poziomie obróbki danych niż z punktu widzenia wyboru cech wypowiedzi.

Poniżej przedstawiono schemat blokowy systemu identyfikacji mówcy, który obejmuje główne zespoły istotnych cech (Marasek et al. 2002 - patent). W świetle nowszych publikacji nie ulega on zasadniczym zmianom, ale można go uzupełnić o wiele elementów, zaś pozostałe uszczegółowić.

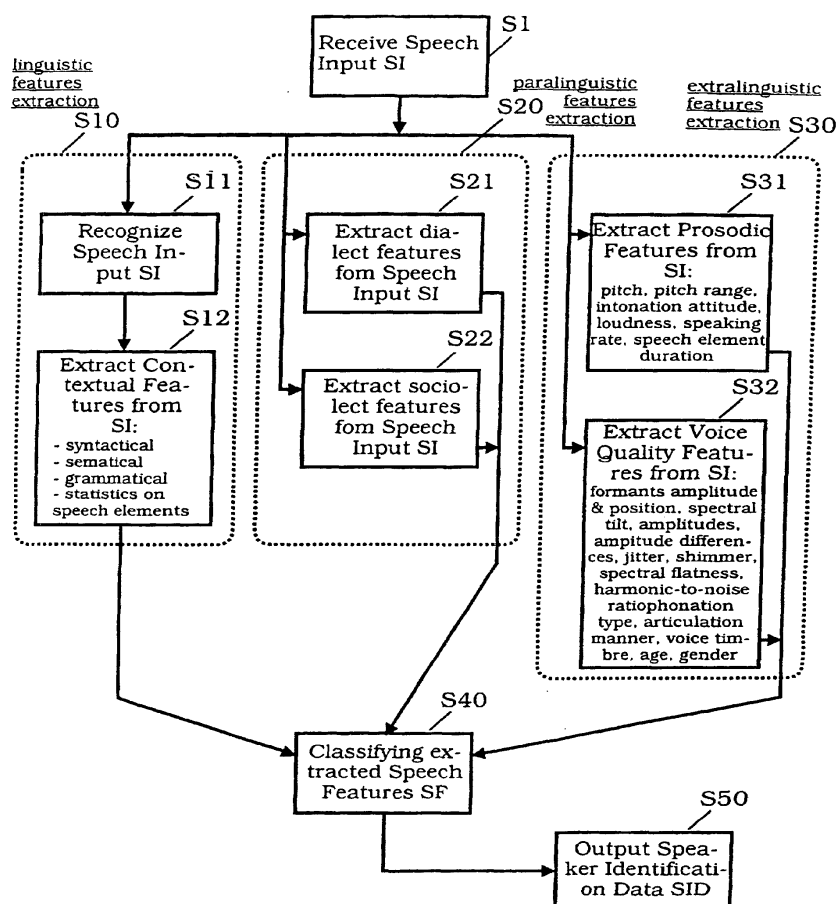


Fig. 1

Ilustracja 1: Inwentarz parametrów akustycznych, cech wypowiedzi i metod statystycznych stosowanych w identyfikacji i weryfikacji mówcy oraz ustalaniu jego cech na podstawie cech sygnału mowy.

Parametr, wielkość, cecha	Zastosowanie, (zakładany) związek z parametrami mówcy, inne uwagi	Przykładowe źródło
„Mierzalne” parametry akustyczne		
<i>Cepstrum, Cepstral Trajectories</i>	speaker identification	Furui 1981
<i>Differential Line Spectral Frequencies</i>	speaker recognition	Ma & Leijon 2011
<i>Formant amplitudes and placement</i>	speaker identification	Becker et al. 2008
<i>Formant frequencies</i>	body size/weight	Gonzalez, J. 2004
<i>Harmonics-to-noise ratio</i>	speaker identification, speaker's age	Ferrand, C. T. 2002
<i>Instantaneous frequencies</i>	speaker identification	Grimmaldi & Cummins 2008
<i>Jitter</i>	speaker identification	Farrus & Hernando 2009
<i>Long-term fo features</i>	speaker's age	Mueller i Burkhardt 2007
<i>LTAS (Long Term Average Spectra)</i>	speaker's age	Linville, S. E. 2002
<i>MFCC + GMM</i>	speaker's age, speaker identification	Chen, Luo 2009; Mahmoodi 2011
<i>Mel-warped cepstrum</i>	speaker identification	Gish & Schmidt 1994
<i>MFCC + HMM</i>	age, speaker identification	Patel, Rao 2010
<i>MFCC + pitch features</i>	gender, speaker identification	DeMarco, Cox 2011
<i>MFCC + SVM</i>	age, speaker identification and verification	Chen & Luo 2009
<i>R-MFCC</i>	speaker identification	Pati, Prasanna 2011
<i>Nasal sounds (as idiosyncratic)</i>	speaker identification	Amino et al. 2006
<i>Pitch frequency</i>	speaker identification	van Dommeln 1987, Carrey 1996
<i>Pitch frequency</i>	attitude, intention	Ishi et al. 2012
<i>Pitch frequency</i>	gender	Mitze et al. 2007
<i>Reflection coefficient</i>	voice identification	Rabiner & Schafer 1989
<i>Rhythm (jako powtarzalność pewnych zjawisk lub ich specyficzny rozkład w</i>	jako wartość mierzalna automatycznie (bez	van Dommelen 1987, Grimmaldi &

czasie)	odniesienia do warstwy językowej wypowiedzi)	Cummins 2010, Dellwo & Koreman 2008
<i>Shimmer</i>	speaker identification	Farrus & Hernando 2009
<i>Short-term Spectral Features</i>	speaker's age	Mueller i Burkhardt 2007
<i>Spectral flatness</i>	speaker identification	Hosseinzadeh, Krishnana 2008
<i>Sub-band energy trajectories</i>	speaker recognition	Adami 2005
<i>Sub-band feature (non-uniform sub-band processing)</i>	ekstrakcja cech fizjologicznych mówcy	Lu & Dang 2007
<i>Supra-segmental level modulation information from residual mel frequency cepstral coefficient (R-MFCC)</i>	speaker identification	Pati & Prasanna 2011
<i>Bispectrum pattern analysis</i>	speaker identification	Fanany, M.I. , Kusumoputro, B. 1998.
<i>Utterance text & text-based/related approaches</i>	ustalenie atrybutów mówcy	Wang et al. 2011, Agrawal et al. 2010
VOT	speaker's age	Decoster, W. and Debruyne, F. 1997
Grupy cech nawiązujące do poziomu językowego wypowiedzi i do jednostek językowych		
<i>text characteristics (całość charakterystyki leksykalnej, składniowej, frazeologicznej)</i>	generalnie	Reynold, Carlson 1995
<i>articulatory characteristics</i>	po części przekłada się na charakter sygnału - jednak nie w 100% (ten sam dźwięk może być wytworzony różną artykulacją)	najczęściej wspominane w tekstach dot. patologii budowy narządu mowy
<i>rhythm, its changes and distortions</i>	w odniesieniu do jednostek językowych (sylaby, akcenty, segmenty fonetyczne)	van Dommelen 1987
<i>speech tempo (measured in linguistic or articulatory units per unit of time)</i>	identyfikacja mówcy, określanie wieku	Quene 2008
<i>pitch contours (kontury intonacyjne - pewne fragmenty przebiegu fo</i>	identyfikacja mówcy, płeć (?)	Atal 1972, Mitze et al. 2007

zintegrowane z jednostkami językowymi, szczególnie z frazą)		
Częściej spotykane algorytmy i metody statystyczne stosowane w identyfikacji i weryfikacji mówcy		
<i>Antimel FCC</i>		Lei 2009
<i>Epoch Vector Strength Comparison</i>	różnicowanie/weryfikacja mówców	Murthy, Yegnanarayana, 2008; Pati, Prasanna 2011
<i>Gaussian Mixture Speaker Models</i>		Reynolds 1995
<i>HMM</i>		bardzo wiele prac odwołuje się do tej metody
<i>i-Vector</i>	speaker recognition	Mandasari et al. 2011
<i>Inner Product Discriminant Function</i>	speaker comparison	Campbell et al. 2009
<i>Linear FCC</i>		Lei 2009
<i>Mel FCC</i>		Lei 2009
<i>Mixed Auto-Associative Neural Networks</i>	speaker verification	Sivaram et al. 2011
<i>Mixture AR HMM</i>	text independent speaker recognition	Tishby 1991
<i>Nearest Neighbour Distance Measure</i>	voice identification	Higgins 1975
<i>Probability Density Estimation</i>		Schwartz et al. 1982
<i>Skew Gaussian Mixture Models for Speaker Recognition</i>	speaker recognition	Matza, Bistriz 2011
<i>Structural Joint Factor Analysis</i>	speaker recognition	Ferras et al. 2011?
<i>Vector Quantization</i>		Soong et al. 1987
<i>Wave Linear Prediction</i>	identyfikacja/weryfikacja	Atal 1974
<i>Super-Dirichlet Mixture Models Using Differential Line Spectral Frequencies</i>	speaker identification	Ma, Leijon 2011
<i>Mixture of Auto-Associative Neural Networks</i>	speaker verification	Sivaram et al. 2011
<i>Eigen-Voice Based Anchor Modeling</i>	speaker identification	Sarkar et al. 2011

1.1.1 Najnowsze badania dotyczące przydatności cech akustycznych do rozpoznawania mowy

Lawson i współpracownicy [2011] dokonują przeglądu cech akustycznych i ich zestawień pod kątem przydatności do rozpoznawania mowy. Poniżej zestawiono pochodzące z tej pracy dane. Autorzy biorą pod uwagę (1) Cechy tradycyjnie stosowane w RM, (2) Nowe zestawy cech, (3) Zmodyfikowane zestawy cech (przedstawione w zamieszczonych poniżej tabelach):

Table 1: Traditional Features for SID

<i>Feature</i>	<i>Coefficients</i>	<i>Freq. Range</i>
MFCC-LL	c1-c19 + Δ	Filter Banks 3-21
MFCC-SRI	c0-c12 + Δ + $\Delta\Delta$ + $\Delta\Delta\Delta$	Filter Banks 3-21
PLP-I4U	c0-c12 + Δ + $\Delta\Delta$	0-4000 Hz
LPCC-I4U	c0-c18 + Δ	0-4000 Hz
LPCC-AFRL	c0-13 + Δ + $\Delta\Delta$	0-4000 Hz

Table 2: Innovative Features for SID

<i>Feature</i>	<i>Coefficients</i>	<i>Freq. Range</i>
PYKFEC (Mel)	0-39 Filters	0-4000 Hz
DPK	c2-c21	0-4000 Hz
MVDR	c1-c12 + Δ	0-4000 Hz
CFC	c1-c30	0-4000 Hz

Table 3: Altered or Extended Feature Configurations for SID

<i>Feature</i>	<i>Coefficients</i>	<i>Freq. Range</i>
LFCC	C1-C19 + Δ	Filter Banks 0-30
Sel LFCC	C1-C19 + Δ	Filter Banks 2-26
Sel LPCC	C0-C13 + Δ + $\Delta\Delta$	300-3400 Hz
Sel PLP	C1-C12 + Δ	300-3400 Hz
Sel PYKFEC (no-Mel)	0-39 Filters	300-3400 Hz

Wyniki testów dla poszczególnych grup parametrów akustycznych zawarto w poniższych trzech tablicach:

Do ewaluacji zastosowano następujące korpusy:



FUNDACJA UNIWERSYTETU
IM. ADAMA MICKIEWICZA
W POZNANIU

ul. Rubież 46, 61-612 Poznań **tel.** 061 827 97 00, fax 061 827 97 01
e-mail: fundacjauam@ppnt.poznan.pl www.fundacja.ppnt.poznan.pl
KRS: 0000052045 **Regon:** 631509229 **NIP:** 781-00-02-075



POZNAŃSKI PARK
NAUKOWO-TECHNOLOGICZNY

Fundacji Uniwersytetu im. A. Mickiewicza

Table 4: Summary of Evaluation Conditions

<i>Condition</i>	<i>Train</i>	<i>Test</i>
Matched Channel Clean 1	Samson C01 Mic Interview	Samson C01 Microphone Interview
Matched Channel Clean 2	Landline Telephone Conversation	Landline Telephone Conversation
Cross Channel Clean 1	Studio Mic Conversation	Landline Telephone Conversation
Cross Channel Clean 2	Landline Telephone Conversation	Studio Mic Conversation
Reverberant Cross Channel	Landline Telephone Conversation	Omnidirectional Mic Conversation
Cross Language	Studio Mic Conversation Native Language	Studio Mic Conversation English
Noisy Matched Channel	Cell phone Public Place	Cell phone Public Place
Noisy Cross Channel	Landline Phone Public Place	Cell phone Public Place
Noisy Cross Channel/Cross Environment	Landline Phone Office	Cell Phone Public Place

Wyniki testów dla poszczególnych grup parametrów akustycznych zawarto w poniższych tablicach:

Table 5: Verification Equal Error Rates for Traditional SID Features

	LPCC- I4U	LPCC- AFRL	PLP- I4U	MFCC- SRI	MFCC- LL
Matched Channel Clean					

Table 6: EER Results for Innovative SID Features

	PYKFEC	DPK	MVDR	CFC
Matched Channel Clean 1	2.27	5.89	2.03	2.45
Matched Channel Clean 2	1.33	19.47	1.57	5.66
Cross Channel Clean 1	8.69	18.73	2.69*	24.79
Cross Channel Clean 2	23.7	24.96	1.49*	15.36
Reverberant Cross Channel	11.57	26.95	10.6	10.08
Cross Language	5.06	12.37	3.24	4.72
Noisy Matched Channel	9.34	13.7	3.42*	9.1
Noisy Cross Channel	13.73	21.25	5.78	12.98
Noisy Cross Channel/ Cross Environment	16.19	23.63	7.02	11.03

Table 7: EER results for Altered or Extended Feature Configurations for SID

	LFCC	Sel LFCC	Sel LPCC	Sel PIP	Sel PYKFEC
Matched Channel Clean 1	1.34	1.06	0.56	1.69	1.12
Matched Channel Clean 2	1.22	1.53	1.19	1.32	0.88
Cross Channel Clean 1	9.13	2.93	5.02	3.85	2.99
Cross Channel Clean 2	7.72	2.31	4.13	4.03	3.9
Reverberant Cross Channel	7.42	7.22	9.2	13.1	10.7
Cross Language	3.57	4.07	5.03	5.29	5.15
Noisy Matched Channel	5.24	6.29	4.42	4.81	6.73
Noisy Cross Channel	5.55*	5.65	5.59	10.5	7.6
Noisy Cross Channel/ Cross Environ- ment	6.72	8.93	8.68	10.9	9.62

1.1.2 Konkluzja

W dostępnych tekstach zorientowanych na zastosowania z jednej strony dominuje nastawienie na badanie krótkoterminowych parametrów akustycznych bez odniesienia do poziomu językowego. Tak więc głos jest traktowany każdy inny sygnał i w zasadzie stosuje się metody takie, jak do identyfikacji innych sygnałów (akustycznych i nie tylko). Jest to zrozumiałe ze względu na możliwość całkowitego zautomatyzowania procesu. Z drugiej strony, takie podejście oznacza rezygnację z możliwości stosunkowo łatwego dostrzeżenia cech z poziomu językowej struktury wypowiedzi (co może przynieść konkretne korzyści).

Wydaje się również, że dyskusja na temat doboru parametrów akustycznych do takich badań jest w ostatnim okresie mało produktywna. Badacze koncentrują się często na samych metodach statystycznych, posługując się bardzo szerokimi zbiorami cech, bez wstępnej refleksji na temat ich wzajemnych wpływów, ich znaczenia, użyteczności, możliwości skutecznej ekstrakcji, itd.

Nowe badania pokazują, że nawet jeśli analiza ma się opierać wyłącznie na parametrach akustycznych, to - o ile jest taka możliwość - warto wykorzystać dłuższe sygnały (nie kilku lecz przynajmniej kilkudziesięciosekundowe) i wprowadzać „miary długoterminowe”. Mogą być to po prostu średnie pewnych parametrów, ale także zakresy ich zmienności, cykliczności, losowości, rozrzut wartości i in.

Wiele wskazuje na to, że badania dotyczące determinacji stanów mówcy na podstawie analizy wypowiedzi są (poza badaniem stanów emocjonalnych) dopiero w początkowej fazie. Badania nad identyfikacją głosu/mówcy są znacznie bardziej zaawansowane.

Wydaje się, że chcąc osiągnąć rezultat zastosowawczy przy obecnym stanie wiedzy, należy rozpocząć od metod „siłowych” wspomaganych wstępną ekspercką analizą sygnałów:

- określać wstępnie charakterystykę badanych sygnałów (włącznie ze sposobem rejestracji, hipotetycznym miejscem zaistnienia, potencjalną grupą mówców, itd.);
- w przypadku różnicowania lub identyfikacji głosów, wybrać optymalny zestaw cech (można tego dokonać na podstawie badań na dostępnych korpusach - badań, które umożliwiłyby określenie optymalnych zbiorów cech dla różnych warunków - zob. poprzedni punkt; można się tutaj również posłużyć wiedzą ekspercką).

Jak wspomniano, treść dostępnych prac wskazuje, że znaczący potencjał tkwi jeszcze na poziomie językowym wypowiedzi, który - ze względu na stopień komplikacji i brak dostępnych modeli - jest rzadko eksploatowany (poza indywidualnymi analizami eksperckimi - tam wykorzystywany jest często, lecz z reguły na poziomie graniczącym z intuicyjnym). W tej sytuacji wymienione dwa

kroki należałoby poszerzyć o kolejne, w których analizowano by wypowiedzi w kategoriach:

- tempo i rytm wypowiedzi i jego zmiany (cykliczność, tendencje, losowość, etc.)
- melodia wypowiedzi, jej wariantywność, typowe przebiegi (j.w.)
- prominencje (realizacja akcentów, ich zróżnicowanie, itd.)
- rozkład zaburzeń płynności (lokalizacje w wypowiedzi, wzajemne lokalizacje)
 - pauzy ciche
 - pauzy wypełnione (również ich realizacja - „zawartość” akustyczna, role w wypowiedzi)
 - inne zawahania, powtórzenia, itd.
- barwa głosu i jej zmiany (w kategoriach subiektywnych i obiektywnych)
- analiza leksykalna wypowiedzi
 - charakter stosowanego słownictwa
 - profile częstościowe
- analiza składniowa
 - typowe konstrukcje składniowe
 - ewentualne powtarzalne błędy
- analiza semantyczna (sposób przekazywania informacji/znaczenia)

Pewna część z wyżej wymienionych „parametrów językowych” jest już przedmiotem analiz pod kątem identyfikacji/rozpoznawania lub budowy profilu charakterologicznego mówcy. Z reguły na podstawie odsłuchu można wstępnie ocenić, które z cech będą odgrywały funkcje różnicujące dla danej grupy sygnałów. Może też o tym decydować grupa potencjalnych mówców (jeśli np. z góry wiadomo, że będą nimi same dzieci). W większości przypadków powyższe analizy związane z poziomem językowym mają rację bytu tylko wówczas, gdy mamy do czynienia z dłuższymi wypowiedziami.

1.2 Propozycje zestawów tagów do anotacji cech paralingwistycznych

Przeprowadzono analizę możliwości kategoryzacji etykiet na potrzeby opisu cech paralingwistycznych. Rozpatrzono analizę cech paralingwistycznych z wykorzystaniem czterech podstawowych wymiarów kategoryzacji: Wymiar I: parametry *lokalne* vs. *globalne*. Ten wymiar może odnosić się do pozostałych wymiarów (np. globalne cechy percepcyjne złożone). Parametry *lokalne* rozumiane jako: (a) dotyczące arbitralnie wybranego/ych fragmentu/ów sygnału (np. 100ms); (b) dotyczące jednostek językowych (np. słowo, sylaba, stopa, fraza, itd.). Parametry *globalne*, dotyczące całego analizowanego sygnału, a więc możliwe do potraktowania jako metadane dla całego sygnału, bez synchronizowania z jego częścią/częściami składowymi. Wymiar II: parametry *bezpośrednie* vs.

złożone/pośrednie (akustyczne). Parametry *bezpośrednie* akustyczne sygnału wynikające z pomiarów bezpośrednich (lub względnie bezpośrednich, zależnych np. od dobranej częstotliwości próbkowania), typu: "punktowego" (np. wartość parametru w danym punkcie czasowym) lub parametry typu "przedziałowego". Parametry *złożone* (akustyczne) - obliczone na podstawie parametrów bezpośrednich (poprzez ich uśrednianie, sumowanie, obliczanie maksimów, sposobów zmienności, etc., parametry typu "przedziałowego" oraz parametry typu "multisample" - mierzone na wybranych próbkach sygnału. Wymiar III: parametry **bezpośrednie vs. złożone/pośrednie (percepcyjne).** Bezpośrednie parametry percepcyjne związane z sygnałem mowy (względnie niezależne od tekstu, np. mieścić oceny emocjonalności, przyjazności, itd. wypowiedzi/brzmienia głosu). Zaliczono tutaj parametry typu "impresyjnego" (np. subiektywna, nieukierunkowana ocena pewnych własności sygnału mowy) oraz typu "odsluchowego" (uzyskane na podstawie laboratoryjnego, ukierunkowanego odsluchu przez eksperta lub grupę ekspertów wg określonej procedury). Złożone (pośrednie) parametry percepcyjne związane z sygnałem mowy - parametry uzyskiwane poprzez operacje dokonane na bezpośrednich parametrach percepcyjnych (np. kombinacje tych parametrów). Wymiar IV: parametry **strukturalne** - związane ze strukturą wypowiedzi na poziomie wyrazu i wyższych, wymagające w analizie odniesienia do struktury językowej: parametry związane ze strukturą językową wypowiedzi, parametry leksykalne (dotyczące cech użytych wyrazów), parametry składniowe (dotyczące operowania składnią, w tym błędów i specyficznych/indywidualnych konstrukcji), parametry dyskursowe (wykraczające ponad pojedyncze frazy - ujawniające się w większych fragmentach tekstu, np. struktura argumentacyjna, często nosząca cechy indywidualne). Część ostatecznie zastosowanych etykiet powinna odnosić się do **parametrów mieszanych**, tj. np. opisujących cechy jednostek językowych trudnych (lub niemożliwych) do automatycznego wyodrębnienia w terminach parametrów fizycznych, np. iloczyn lub energia sylaby, wyrazu, frazy.

Z myślą o implementacji w opracowywanym systemie anotacji nagrań mowy opracowano zestawy różnego typu tagów. Oparto je na:

1. istniejących technikach i metodach opisu mowy (np. GRBAS)
2. wybranych podejściach do opisu cech parajęzykowych wypowiedzi oraz cech mówcy, nie oferujących jednak własnego systemu opisu tych cech;
3. badaniach własnych.

Wzięto również pod uwagę tagset oparty na kategoryzacji aktów dialogowych DIT++, gdyż posługiwanie się kategoriami aktów dialogowych może się okazać niezbędne w analizie funkcji parajęzykowych cech wypowiedzi.

Tagsety:

VQLaver - ocena barwy głosu wg Lavera (1980) (konfiguracja laryngalna i supralaryngalna - sześć wymiarów, tagset raczej dla specjalistów fonetyków)

EASE - Emotional Attitude Subjective Estimation - subiektywna ocena emocji nastawienia mówcy (dla „naiwnych” anotatorów, prosta lista cech z możliwością określenia ich natężenia).

GRBAS - Grade, Roughness, Breathiness, Asthenia, Strain Scale (voice disorder assessment) (Isshiki & Takeuchi 1970; Hirano 1981, 1989; system popularny w USA i w Japonii; krytykowany za słabą ortogonalność wymiarów)

RBH - Roughness, Breathiness, Hoarseness (voice disorder assessment) - prosty tagset do oceny patologii głosu (powstały na bazie krytyki GRBAS (Wendler et al. 1986))

VQPA - percepcyjna ocena barwy (normalnego) głosu. Tagset „uniwersalny”, duża liczba tagów, dla naiwnych anotatorów i specjalistów - można wybrać subsety; część tagów wykorzystuje „suwak”, część to tagi „proste” (binarne - obecność/brak cechy).

SCSA - Speech Clarity Subjective Assessment subiektywna ocena wyrazistości i zrozumiałości wypowiedzi

GSPA - General Subjective Prosody Assessment - zestaw „ogólnych” miar prozodycznych do ocen percepcyjnych.

DADIT++ - akty dialogowe na podstawie systemu DIT++

1.3 Właściwości akustyczne pauz wypełnionych w mowie spontanicznej

Przeprowadzono wstępne badanie akustycznych właściwości pauz wypełnionych w mowie spontanicznej. Zdecydowano się na badanie nagrań dialogowych, ze względu na ich specyficzne cechy, odmienne niż te które cechują monolog (częściej dostępne w laboratoryjnych bazach danych), z pewnością można oczekiwać większej liczby nieciągłości, chociażby ze względu na (zupełnie lub częściowo) nieprzewidywalny wpływ interlokutora na konstruowane wypowiedzi. Przedmiotowe pauzy wypełnione anotowano manualnie, dokonano automatycznej ekstrakcji z nagrań dialogowych (20 dialogów), a następnie przeprowadzono pomiarów parametrów akustycznych dotyczących melodii (częstotliwość podstawowa i jej zmiany) oraz jakości głosu tzw. „voice quality” (wartości parametrów *jitter*, *shimmer*, a także częstotliwości formantowe). Zestawiono wyniki pomiarów porównano dla wszystkich mówców oraz różnych grup mówców. Homogeniczne, jednorodne pauzy wypełnione są względnie łatwe do identyfikacji w materiale nagraniowym i jest możliwa ich automatyczna analiza.

Wiele z nich zachowuje względnie stałe wartości częstotliwości podstawowej oraz parametrów „voice quality”. Jednakże wciąż nie jest pewne, na ile różnice tych właśnie wartości mogą wspomagać rozróżnianie, identyfikację czy rozpoznawanie mówcy.

1.4 Analiza zawartości bazy Jurisdict pod kątem badania cech paralingwistycznych

Sposób wykonania przeglądu

1. Zapoznano się z dokumentacją dotyczącą struktury i zawartości bazy.
2. Przeanalizowano scenariusze sesji nagranych i wybrano z nich wypowiedzi potencjalnie przydatne do dalszych badań. Były to przede wszystkim wieloczęściowe wypowiedzi quasi-spontaniczne (narracja: opis wypadku, wykroczeń w ruchu drogowym, drogi do domu, itp.), wypowiedzi realizowane jako „odpowiedzi na pytania”, „polecenia”, oraz niektóre próbki wypowiedzi czytanych.
3. Odsłuchano realizacje tych wypowiedzi u około 180 mówców z sekcji „police” (pierwszych 180 sesji spełniających kryteria).
4. Kilka wrywkowo wybranych sesji odsłuchano niemal w całości.
5. We wczesnej fazie odsłuchu (po przesłuchaniu ok. 20 mówców) stwierdzono, że „odpowiedzi na pytania” oraz quasi-spontaniczne „polecenia” nie posiadają oczekiwanych własności i zrezygnowano z ich dalszego odsłuchiwanie.
6. W pozostałych głosach skupiono się ostatecznie na wypowiedziach o charakterze „narracyjnym” (p. 2).
7. Wstępnie odsłuchano wybrane fragmenty kilku sesji pochodzących z różnych rejonów kraju pod kątem regionalizmów.

Podsumowanie wyników odsłuchu

Materiał jest bardzo bogaty i umożliwia przeprowadzenie wielu analiz, lecz z punktu widzenia analizy cech para- i pozajęzykowych posiada on również istotne wady:

1. Wypowiedzi nie były realizowane jako autentyczne akty komunikacyjne: nie były kierowane do konkretnego odbiorcy i w konkretnym celu (poza samym celem rejestracji). Na tej podstawie można przypuszczać, że pod wieloma względami będą odbiegały charakterem od wypowiedzi tych samych mówców w warunkach komunikacji (nawet formalnych sytuacji komunikacyjnych). W warstwie prozodycznej wypowiedzi są często „martwe”: nie ma typowej dla mowy spontanicznej zmienności melodycznej i rytmicznej, są za to pewne stałe tendencje, które pozwalają zidentyfikować wypowiedzi jako „sztuczne”: powtarzane monotonne przebiegi melodyczno-rytmiczne.
2. Wypowiedzi rzadko mają charakter spontaniczny w sensie konieczności sformułowania od podstaw przez mówcę w reakcji na określoną sytuację komunikacyjną. Tutaj nawet wypowiedzi nie mające znamion

„szablonowych” (frazologia profesjolektu), często są realizowane w sposób zdeterminowany przez technikę i metody rejestracji.

3. Wypowiedzi mają charakter quasi-formalny lub formalny i bardzo specyficzny styl (wynikający z przeznaczenia bazy), który uniemożliwia generalizację ewentualnych spostrzeżeń słownictwa, frazeologii, składni, itd.. Mówcy starają się posługiwać językiem urzędowym, profesjolektem typowym dla raportów policyjnych. Ze względu na sposób realizacji nagrań, nawet dłuższe wypowiedzi są „sztucznie” dzielone na frazy lub sekwencje fraz przez samego mówcę - również w obrębie jednego sygnału: Pojawiają się długie pauzy między frazami, brakuje ponadfrazowej spójności melodycznej (z wyjątkiem tego, że słysząc, że mówi to ta sama osoba). Prawdopodobnie wypowiedziane bez pauz (wymuszonych przez procedurę lub narzucanych sobie przez mówcę) brzmiałyby inaczej. Takie dzielenie często pociąga za sobą monotonię rytmiczną i intonacyjną. Kolejne składowe wypowiedzi są realizowane na zasadzie „wyliczania”.

Analizy możliwe do przeprowadzenia na materiale w obecnej postaci

Na dostępnym materiale można przeprowadzić pewne analizy dotyczące cech widmowych wypowiedzi, części cech segmentalnych (wymowa - np. w kategoriach procesów lenicyjnych i fortycyjnych). Można ustalić specyfikę wymowy poszczególnych mówców na poziomie segmentalnym (realizacja pojedynczych segmentów i ich sekwencji - np. zbitek). W pewnym zakresie możliwe jest określenie różnic w charakterze wymowy tekstu czytanego i mówionego bez przygotowania. W ograniczonym zakresie można analizować cechy suprasegmentalne (zmienność *fo*, realizacja i rozkład prominencji, itd.) oraz realizację składniową i frazeologiczną wypowiedzi (tj. przede wszystkim jej usterki).

Dłuższe wypowiedzi o przynajmniej minimalnym poziomie „spontanizacji” zawierają stosunkowo dużo elementów charakterystycznych, które można klasyfikować jako „parajęzykowe”, a w szczególności wszelkiego rodzaju zawahania zaburzenia płynności wypowiedzi:

- pauzy ciche
- pauzy wypełnione
- wydłużenia samogłosek i spółgłosek (o charakterze „zawahań”)
- wtrącenia (nieliczne)

Jednak duża część tych zjawisk występuje w szczególnych miejscach wypowiedzi - często na początku. Można przypuszczać, że w mowie spontanicznej dystrybucja i lokalizacja tych zjawisk będzie jednak inna.

W wypowiedziach „quasi-spontanicznych” (np. opis wykroczeń drogowych) często pojawiają się różnego rodzaju błędy gramatyczne, logiczne, i in. - wydaje się, że ich charakter często jest specyficzny i może stanowić podstawę identyfikacji mówcy.

Odpowiedzi na pytania to zwykle bardzo krótkie frazy, często pojedyncze wyrazy. Jednak w nagraniach słychać, że były prawdopodobnie realizowane jedno po drugim. Niekiedy całe sekwencje odpowiedzi danego mówcy są realizowane na bazie tej samej melodii. Materiał ten nie spełnia więc oczekiwań zarówno pod względem długości wypowiedzi, jej zawartości, jak i sposobu realizacji.

Zdania „bogate fonetycznie” też posiadają pewną wartość. W przypadkach trudniejszych artykulacyjnie sekwencji widać, jakie tendencje są powszechniejsze, jakie zaś charakterystyczne dla niektórych mówców.

Wypowiedzi czytane również można wykorzystać. Pozwalają badać charakterystykę sposobu czytania tekstu (np. skłonność do hiperpoprawności), chociaż niektóre jej elementy byłyby możliwe do zbadania dopiero w przypadku dłuższych czytanych tekstów ciągłych.

Wprawdzie dokonano jedynie bardzo wybiórczego odsłuchu niewielkiej liczby wypowiedzi pod kątem regionalizmów, ale nie dostrzeżono ich wyraźnej obecności (na poziomie fonetycznym). Warto jednak ten aspekt zbadać głębiej, bo baza mogłaby stać się źródłem danych do identyfikacji pochodzenia mówców. Byłoby to możliwe, gdyby dostępne były dane o miejscach nie tylko urodzin, ale również dłuższego pobytu mówców w różnych okresach życia.

Wskazane uzupełnienia materiału

Jak podkreślano, na dostępnym materiale można wykonać szereg analiz, lecz ich wyniki mogą być trudne lub niemożliwe do wykorzystania w odniesieniu do szerszej palety typów wypowiedzi.

Wskazane byłoby pozyskanie materiału, który obejmowałby wypowiedzi:

- znacznie dłuższe, ciągłe;
- o charakterze (quasi)spontanicznym;
- o charakterze interakcyjnym (np. dialogowe);
- wypowiedzi realizowane w warunkach stresu lub możliwych do zidentyfikowania stanach emocjonalnych.

Tego rodzaju materiał pozwoliłby na przeprowadzenie większej liczby analiz oraz na uogólnianie wyników przynajmniej poza grupę zawodową i poza sytuację „nagrań

studyjnych”. Nawet wstępny odsłuch dłuższych wypowiedzi mógłby prawdopodobnie naprowadzić na różne typy idiosynkratycznych cech wymowy, które być może warto byłoby przełożyć na parametry i potraktować jako potencjalne wskazówki w identyfikacji głosu.

Możliwe byłyby m.in. następujące badania, obejmujące wiele podsystemów języka:

- analiza długoterminowej zmienności parametrów widmowych (szczególnie w wypadkach wypowiedzi emocjonalnych, ale również pod kątem zmęczenia głosu; wiarygodne określenie barwy głosu i jej typowych zmian, itd.);
- rozkład zaburzeń płynności (jako cecha idiosynkratyczna, również w powiązaniu z budową wypowiedzi na poziomie słownictwa i składni, itd.);
- charakterystyka rytmu wypowiedzi oraz jego zmienności lub ewentualnych zaburzeń;
- charakterystyka i powtarzalność konturów intonacyjnych dla określonych kategorii wypowiedzi (aktów mowy);
- zmienność *fo* w dłuższych przedziałach czasowych, zanikanie *fo*, zakres ubezdźwięczniania;
- charakterystyka frazy (przeciętna długość, budowa);
- specyfika („profil”) wypowiedzi na poziomie leksykalnym, frazeologicznym i składniowym.

Pozyskanie nagrań mowy emocjonalnej lub choćby „w pełni spontanicznej” jest niezwykle trudne. W warunkach laboratoryjnych można próbować np. następujących technik:

- nagrania z różnymi „dystraktorami” (bodźce zaburzające uwagę, np. konieczność wykonywania jakiejś dodatkowej czynności, pokazujące się obrazy, dźwięki, itd.);
- trudne pytania lub obciążające poznawczo zadania, wprawiające rozmówcę w zakłopotanie albo trudne merytorycznie, prowadzące do sytuacji, w której koncentruje się na zadaniu/zagadnieniu, a mniej kontroluje mowę;
- nagrania rozmów w większej grupie (większe prawdopodobieństwo oderwania uwagi od faktu bycia nagrywanym, większa szansa na rejestrację zachowań spontanicznych);
- realizacja zadań w parze lub większej grupie.

Jednak z dostępnych prac wiadomo, że w takich warunkach paleta stanów mówcy jest stosunkowo spłaszczona i ograniczona. Nielaboratoryjny sposób pozyskania danych to np. realizacja nagrań techniką podobną do stosowanej w rejestracji korpusu Cambpella (nagrywarki stale noszone przez osoby nagrywane), nagrania ze spotkań, zebrań, dyskusji. Oczywiście, takie rodzaje nagrań wnoszą pewne ograniczenia techniczne.

Konkluzja

- Na dostępnym materiale można wykonać wiele rodzajów analiz, w tym ekstrakcji niektórych parametrów akustycznych głosu i prób powiązania ich np. z wiekiem, płcią i wagą (?) mówcy. Wydaje się, że materiał ten może być wystarczający do testowania niektórych metod identyfikacji głosu/mówcy, ale posiada on też szereg ograniczeń związanych ze sp
- specyfiką mówców i charakterem zarejestrowanych wypowiedzi. Informacje o mówcach mogą być niewystarczające.
- Efektywne wykorzystanie analizowanej części bazy do badań zjawisk para- i pozajęzykowych w wypowiedziach wymagałoby przetranskrybowania jej zawartości na poziomie przynajmniej fonologicznym i otagowania zjawisk para- i pozajęzykowych.
- Do analiz wielu cech parajęzykowych potrzebne są dłuższe, przynajmniej kilku-kilkunastofrazowe ciągłe wypowiedzi, których liczba tutaj jest ograniczona. Są one pod wieloma względami specyficzne, co może spowodować, że wyniki ewentualnych analiz trudno będzie generalizować.
- Potrzebne są również nagrania o wyższym poziomie spontaniczności - spontaniczne wypowiedzi w rozmowie lub w ramach zadań, które odwracałyby uwagę od samego formułowania i artykułowania wypowiedzi.
- Ważny jest też stopień interaktywności wypowiedzi (powiązany zwykle z jej spontanicznością).

1.5 Analiza porównawcza cech indywidualnych wypowiedzi w mowie natywnej i nienatywnej

Przeprowadzono analizę realizacji struktur rytmicznych w mowie natywnej i nienatywnej, celem zdefiniowania miar akustycznych dla badania indywidualnych strategii stosowanych przez mówców posługujących się językiem polskim jako językiem ojczystym i obcym.

Jednym ze spodziewanych rezultatów była obserwacja i klasyfikacja zjawisk związanych z przeniesieniem wzorców czasowych z języka rodzimego i zastosowaniem ich w drugim języku. Tego rodzaju informacja byłaby użyteczna m.in. w analizie i ocenie obcego akcentu w mowie nienatywnej oraz w ogóle w (także automatycznym) odróżnieniu mówców natywnych od nienatywnych. W badaniach użyto materiałów nagraniowych, pochodzących od mówców, których pierwszym językiem był język polski, niemiecki oraz koreański. Języki te reprezentują pod względem organizacji rytmicznej wypowiedzi różne rodziny językowe: status języka polskiego jest różnie interpretowany w dotychczasowych badaniach w dziedzinie przedmiotu (jako oparty na sylabie, oparty na akcencie lub też pośredni), język niemiecki stanowi typowy przykład języka o rytmie opartym na akcencie. W odniesieniu do języka koreańskiego istnieją rozbieżności w jego

klasyfikacji rytmicznej: niektórzy autorzy wskazują, że rytm w koreańskim oparty jest raczej na sylabie niż na akcencie, ale istnieją również badania wskazujące na morę jako jednostkę rytmu w tym języku.

Wyniki przeprowadzonych badań pokazały, że zastosowane parametry akustyczne (%V, ΔC , nPVI oraz rPVI) skutecznie pozwalają na uchwycenie różnic w realizacji rytmu w polskiej mowie natywnej i nienatywnej, i że różnice te spowodowane są w głównej mierze inną realizacją czasową interwałów samogłoskowych (opisywanych przez parametry %V i nPVI). Stwierdzono również, że %V i ΔC (wyrażające kolejno zmienność w czasie trwania interwałów samogłoskowych i spółgłoskowych) są znacznie bardziej efektywne w kodowaniu różnic w realizacji rytmu w mowie natywnej i nienatywnej oraz między mówcami posługującymi się różnymi L1 (językami ojczystymi) niż parametry PVI. Na podstawie rozkładu wartości wyżej wymienionych miar oraz wyników analizy statystycznej (ANOVA i porównania post-hoc, test Tukey'a) stwierdzono, że największa różnica pod względem realizacji rytmu w polskiej mowie nienatywnej dotyczy mówców Koreańskich, którą to obserwację potwierdziły również odsłuchy nagrań. Ponadto, parametr %V może być z powodzeniem stosowany w analizie indywidualnej realizacji rytmu przez mówców (wykazano statystycznie istotne różnice w jego wartościach między mówcami, niezależnie od L1). Zastosowane metody pomiaru rytmu potwierdziły do pewnego stopnia hipotezę dotyczącą transferu realizacji rytmu z języka ojczystego do języka obcego - dotyczyło to mówców Koreańskich i sposobu, w jaki realizowali oni czasowo (rytmicznie) interwały samogłoskowe. Jednak, co istotniejsze, wartości %V, ΔC i parametrów PVI odzwierciedlają różnice w rytmie mowy, wynikające z odmiennych struktur fonotaktycznych w badanych językach: wszyscy mówcy nienatywni, niezależnie od L1, wykazali znacznie większe zróżnicowanie w realizacji czasowej interwałów spółgłoskowych niż mówcy polscy. Wskazuje to na fakt, że realizacja spółgłosek była dla nich szczególnie trudna, co ma swoje źródło w strukturze polskiej sylaby, która jest uznawana za wyjątkowo złożoną. W odróżnieniu od języka niemieckiego czy koreańskiego, w języku polskim wieloelementowe zbitki spółgłoskowe pojawiają się bardzo często i to zarówno w onsecie jak i codzie sylaby (lub w obu naraz). Stanowi to wyjątkową trudność dla osób posługujących się językiem polskim jako obcym i prowadzi do odmiennej realizacji czasowej interwałów spółgłoskowych.

2 Kontynuacja prac nad analizą emocji

Modelowanie i rozpoznawanie emocji na podstawie cech głosu mówcy

2.1 Cel eksperymentu

Celem eksperymentu jest określenie profilu akustycznego charakteryzującego mowę nacechowaną wybranymi emocjami (z grupy tzw. emocji podstawowych). W tym celu, korzystając z istniejącej bazy mowy emocjonalnej (będącej własnością Zakładu Elektroniki Medycznej Politechniki Łódzkiej) przeprowadzone zostaną test percepcyjny oraz analiza akustyczna. Ich wyniki zostaną wykorzystane do budowy modelu statystycznego pozwalającego na rozpoznawanie emocji na podstawie cech głosu mówcy.

2.2 Wstępne testy z wykorzystaniem Emocjonalnej bazy Danych Politechniki Łódzkiej

Mówcy

Baza zawiera 240 wypowiedzi zrealizowanych przez 8 mówców (4 kobiety, 4 mężczyzn). Mówcy to profesjonalni aktorzy, absolwenci Państwowej Wyższej Szkoły Filmowej, Telewizyjnej i Teatralnej w Łodzi:
(http://www.eletel.p.lodz.pl/bronakowski/med_catalog/).

Materiał językowy

Każdy mówca przeczytał 5 neutralnych leksykalnie i znaczeniowo zdań:

Oni kupili dzisiaj nowy samochód. / Jego dziewczyna przylatuje dzisiaj samolotem.

Janek był dzisiaj u fryzjera. / Ta lampa dzisiaj jest na biurku. / Od dzisiaj przestaje się golić.

Emocje

Każde zdanie zostało zrealizowane w sposób neutralny, pozbawiony nacechowania emocjonalnego (ekspresji) oraz w każdym z pięciu stanów emocjonalnych: *radość*, *smutek*, *strach*, *złość*, *znudzenie*.

Emocje te zaliczane są do emocji podstawowych i reprezentują różne rodziny emocji (Ekman 1992) i pozwalają na zbadanie czynników biorących udział w produkcji i percepcji mowy nacechowanej emocjami o odmiennym stopniu

pobudzenia-aktywacji (tj. innej intensywności) oraz odmiennej jakości (lub wartościowości, tj. emocje pozytywne - negatywne).

W celu wzbudzenia określonego stanu emocjonalnego zastosowano metodologię zaproponowaną w pracy Mozziconacci & Hermesa (1999): realizacja zdania z konkretnym nastawieniem emocjonalnym poprzedzona była realizacją wypowiedzi, której warstwa leksykalna była nośnikiem pożądanego nacechowania emocjonalnego.

Nagrania

Miejszem realizacji nagrań była Państwowa Wyższa Szkoła Filmowa, Telewizyjna i Teatralna w Łodzi. Nagrań mówców dokonywano w trakcie jednej sesji wykorzystując mikrofon pojemnościowy; format nagrań: 44,1 kHz, 16 bitów.

Wstępna walidacja nagrań

Realizacja emocji została poddana walidacji w teście percepcyjnym, w którym wzięło udział 50 osób. Wybrano losowo 60 wypowiedzi (10 dla każdego stanu emocjonalnego) i poproszono słuchaczy o sklasyfikowanie ich pod względem emocji (należało wskazać jedną z emocji lub stan neutralny). Średnia dokładność klasyfikacji wyniosła 72%, przy czym u różnych mówców wahała się ona między 60 a 84%.

Anotacja

Cały materiał został automatycznie przetranskrybowany i posegmentowany na poziomie głosek. Transkrypcja i segmentacja zostały ręcznie poprawione.

Baza polskiej mowy emocjonalnej spełnia wymogi stawiane tego rodzaju bazom w zakresie liczby mówców, reprezentowanych emocji, materiału językowego. Podobne bazy mowy emocjonalnej są szeroko wykorzystywane w analizie i modelowaniu emocji, m.in. Hammerschmidt & Jürgens (2007), Saratzaga et al. (2006), Burkhardt et al. (2005), Ververidis et al. (2004), Liscombe et al. (2003), Montero et al. (1999), Montero et al. (1999), Banse & Scherer (1996).

Wyczerpujący przegląd wszystkich dostępnych baz mowy emocjonalnej przedstawiono w pracy „Emotional speech recognition: Resources, features, and methods” (Ververidis & Kotropoulos).

2.2.1 Konkluzja

W trakcie wstępnych analiz stwierdzono niezadowalającą jakość techniczną badanego materiału dźwiękowego w bazie Politechniki Łódzkiej, znaczną ilość szumów. Ponadto w bazie reprezentowany był stosunkowo niewielki zbiór emocji (5). W związku z powyższym, w ramach obecnego projektu zrezygnowano z dalszych prac z wykorzystaniem ww. bazy mowy emocjonalnej i podjęto decyzję o poszerzeniu bazy danych „Paralingua” o wysokiej jakości nagrania mowy emocjonalnej (zob. niżej, rozdz. 3).

2.3 Analiza percepcyjna

Nagrania zostaną poddane bardziej szczegółowej analizie percepcyjnej aby umożliwić wyłonienie w analizie akustycznej tych cech głosu, które słuchacze biorą pod uwagę wnioskując o stanie emocjonalnym mówcy.

Zadaniem słuchaczy (10 osób) będzie sklasyfikowanie stanu emocjonalnego mówcy na podstawie cech jego głosu podczas realizacji każdej z 5 wypowiedzi w każdym z 5 stanów emocjonalnych, plus mowa neutralna. Mając do dyspozycji tylko głos mówcy słuchacz wskaże, czy dana wypowiedź jest emocjonalnie neutralna lub nacechowana strachem, radością, smutkiem, złością lub znudzeniem.

Wyniki testu percepcyjnego posłużą do określenia średniej dokładności z jaką słuchacze na podstawie cech głosu są w stanie określić stan emocjonalny mówcy. Wyniki zostaną również dokładnie przeanalizowane pod kątem dokładności, z jaką rozpoznawane są poszczególne emocje: wyłonione zostaną emocje najbardziej i najmniej rozpoznawalne, co pozwoli na sformułowanie hipotez dotyczących wzorców akustycznych charakteryzujących ekspresję głosową tych emocji.

2.4 Analiza akustyczna

Choć do tej pory nie udało się jednoznacznie zidentyfikować akustycznych wzorców charakteryzujących ekspresję głosową różnych emocji to wiadomo, że udział w niej biorą następujące parametry akustyczne: *parametry źródła głosu* w tym ogólna jakość głosu oraz parametry prozodyczne, *parametry artykulacyjne* oraz *tempo i płynność mowy* (Scherer 2003).

Związek między konkretnymi zmiennymi akustycznymi a emocjami przedstawia się następująco:

Złość: wzrost średniego F0 i średniej energii, większy zakres i zmienność F0 (w przypadku „hot anger”, ale nie „cold anger”). Ponadto wypowiedzi nacechowane gwałtowną złością cechuje także wzrost energii w wyższym paśmie częstotliwości, szybsza artykulacja oraz kontury opadające (downward-directed F0 contours).

Strach: podobnie jak złość charakteryzuje się wysokim stopniem pobudzenia, a więc można się spodziewać wzrostu średniego F0 i jego zakresu, wysokiego poziomu energii w wyższym paśmie częstotliwości (-> tense voice) oraz przyspieszonej artykulacji.

Smutek: jest emocją o niskim stopniu aktywacji, więc można spodziewać się spadków zamiast wzrostów. Dotyczy to średniego F0 i jego zakresu, średniej energii, niższego poziomu energii w wyższym paśmie częstotliwości (-> lax voice) i wolniejszej artykulacji oraz konturów opadających (downward-directed F0 contours).

Jednak intensywniejsze formy smutku, np. rozpacz, mogą być kodowane za pomocą wzrostu średniego F0 i energii.

Radość: podobnie jak w przypadku emocji o wysokim stopniu pobudzenia można się spodziewać wzrostu średniego F0, jego zakresu i zmienności, średniej intensywności, a także podwyższonej energii w wyższym paśmie częstotliwości (-> tense voice) i szybszej artykulacji.

Znużenie: spadek średniego F0 i intensywności, niektóre badania wykazały wzrost, inne spadek tempa mowy, co może świadczyć o braku wyraźnego wzorca akustycznego dla ekspresji tej emocji (jest to o tyle prawdopodobne, że emocja ta nie ma tak silnego podłoża biologicznego jak np. strach).

Wyniki Scherera będą potraktowane jako hipotezy, które w obecnym badaniu będą testowane.

Dla każdej z 280 wypowiedzi zostaną określone następujące parametry akustyczne:

- częstotliwość podstawowa: średnia (-> F0 level), odchylenie standardowe (-> F0 range), 25 i 75 percentyl F0
- energia (intensywność): średnia
- tempo mowy: czas trwania fragmentów wypowiedzi, w których nie zarejestrowano ciszy (non-silent periods - articulation periods), czas trwania dźwięcznych fragmentów wypowiedzi; krótszy iloczyn będzie oznaczał szybsze tempo wypowiedzi
- widmo długookresowe fragmentów dźwięcznych: proporcja energii w wybranych pasmach częstotliwości, różnica między maksimum energii w paśmie 0-2000Hz i 2000-5000Hz, spadek energii powyżej 1000Hz, stosunek energii w paśmie do 500Hz i 1000Hz (-> miara względnego poziomu energii w wysokim paśmie częstotliwości wzgl. niskiego)
- widmo długookresowe fragmentów bezdźwięcznych: proporcja energii w wybranych pasmach częstotliwości

W celu zbadania związku między badanymi parametrami a zmiennymi: zdanie (5), płeć mówcy (2), mówca (8), emocja (6) przeprowadzona zostanie analiza

wielokrotnej regresji liniowej. Można spodziewać się wpływu zmiennych (w różnym zakresie) na badane parametry, szczególnie zaś te związane z F0 (głównie z uwagi na płeć mówców), jednak wpływ „emocji” powinien być dominujący i możemy oczekiwać, że ta zmienna będzie wyjaśniała w największym stopniu różnice w wartościach parametrów. Aby dokładnie to ustalić, trzeba będzie znormalizować wartości parametrów akustycznych względem tych zmiennych, które istotnie wpływają na wartości parametrów (z pewnością będzie to mowa).

Znormalizowane wyniki posłużą do zbadania korelacji między parametrami, aby na dalszym etapie badania można było wykluczyć te parametry, które wnoszą najmniej informacji na temat stanu emocjonalnego mówców. Wyniki regresji pozwolą również na określenie udziału poszczególnych parametrów akustycznych we wzorcach ekspresji głosowej poszczególnych emocji.

Następnie zostanie przeprowadzona kolejna *wielokrotna regresja liniowa*, która da odpowiedź na pytanie, czy wzorce akustyczne ekspresji emocji zawierają te cechy, którymi słuchacze kierowali się w percepcyjnej ocenie emocji. W analizie soatną uwzględnione tylko te parametry akustyczne, które w największym stopniu przyczyniają się do zróżnicowania wypowiedzi emocjonalnych (zbiór wyłoniiony na podstawie wyników wcześniejszej regresji liniowej i analizy korelacji).

2.5 Rozpoznawanie emocji - modelowanie statystyczne

Wykorzystana zostanie *analiza dyskryminacyjna* oraz *sieci neuronowe*. W przypadku rozpoznawania emocji na podstawie cech akustycznych dokładność zaprojektowanych modeli będzie tak samo istotnym wynikiem jak wzorce błędów klasyfikacyjnych. Porównanie klasyfikacji dokonanej przez słuchaczy (w teście percepcyjnym) oraz przez modele pokaże na ile wyniki są zbieżne, czy słuchacze i modele polegali na tych samych, czy może innych wzorcach akustycznych wnioskując o stanie emocjonalnym mówcy.

3 Korpus nagrań „Paralingua”

3.1 Struktura korpusu

Korpus „Paralingua” składa się z dwóch głównych podzbiorów: „Dial” (spontaniczne dialogi zadaniowe) oraz „Emo” (mowa nacechowana emocjonalnie). Dodatkowo jako odniesienie oraz w celu umożliwienia analiz porównawczych każdy mówca, którego głos zarejestrowano w obu podzbiórach przeczytał dwukrotnie ten sam krótki tekst („Północny wiatr i słońce”, Bajka Ezopa).

3.2 Dialogi zadaniowe - podzbiór „Dial”

Opracowano procedurę nagraniową mającą na celu pozyskanie spontanicznych i quasi-spontanicznych nagrań dialogowych. Procedura zakłada nagrania audio i video rozmów par mówców, prowadzących dialog przez telefony komórkowe. Każda z par

otrzymuje polecenie, aby wspólnie zrealizowała cztery zadania dialogowe. Ponadto każdy z mówców indywidualnie czyta krótki tekst. Rozmówcy przebywają w oddzielnych pomieszczeniach, w trakcie trwania sesji nagraniowej ich jedyną możliwością kontaktu jest wymiana informacji za pośrednictwem telefonu. Zadania realizowane w parach miały na celu przede wszystkim poszerzenie dostępnego materiału nagraniowego o dane dialogowe, o większym poziomie spontaniczności i innej charakterystyce zawartości cech paralingwistycznych. Tematykę zadań dialogowych dobrano w taki sposób, aby dodatkowo uzyskać realizację terminologii i struktur związanych m.in. z pozyskiwaniem i przetwarzaniem informacji słownych w systemach zapobiegania oraz zwalczania przestępczości i terroryzmu: podanie opisów miejsc (uzyskanie/dostarczenie informacji o wyglądzie budynku, pomieszczenia, przedmiotów) podanie rysopisu osoby (uzyskanie/dostarczenie informacji, identyfikacja osoby). W trakcie trwania rozmowy (po zakończeniu realizacji pierwszego zadania), mówcy są informowani o tym, że w realizacji kolejnych zadań istotną rolę odgrywa czas.

Zgodnie z ww. procedurą przeprowadzono nagrania, w ramach których zarejestrowano łącznie 30 mówców w przedziale wiekowym 20-35 lat. W rezultacie otrzymano zbiór nagrań obejmujący następujące kanały dla każdej sesji nagraniowej: dźwięk z mikrofonu nagłownego (dla obojga rozmówców z każdej z par), dźwięk zarejestrowany bezpośrednio za pomocą telefonów komórkowych (dla obojga rozmówców z każdej par), dźwięk z telefonu komórkowego zarejestrowany przez komputer (dla jednego z rozmówców), obraz wideo z kamery wraz z dźwiękiem (dla obojga rozmówców z każdej z par).

3.2.1 Procedura nagraniowa

3.2.2 Założenia wyjściowe i warunki nagrań

1. Nagrywana jest rozmowa dwóch osób, rozmawiających przez telefon
2. Mówcy przebywają w dwóch osobnych pomieszczeniach. Jedno z pomieszczeń to pokój biurowy Zakładu Fonetyki UAM, drugie, to studio nagrań, znajdujące się w budynku Collegium Novum UAM, udostępnione na potrzeby nagrań przez *Center for Speech and Language Processing* działającym przy UAM. Użycie tych pomieszczeń było korzystne także ze względów praktycznych, osoby nagrywane to studenci i pracownicy przebywający na co dzień w budynku Collegium Novum UAM.
3. Sprzęt i oprogramowanie:
 - program Sony Sound Forge Audio Studio ver. 10.0
 - Oprogramowanie Record My Call 4.11 na telefony komórkowe, do nagrywania rozmów telefonicznych (dla systemu Android)
 - 2 telefony komórkowe z zestawami słuchawkowymi; zastosowane telefony komórkowe: LG GT540 Swift (telefon w studio) oraz Samsung Galaxy Ace GT-S5830 (biuro).

- słuchawki (dodana przejściówka ze słuchawek na wejście komputera)
- dwie kamery BENQ DVM23 (Full HD 1080P, 5x Optical Zoom)
- laptopy oraz mikrofony nagłowne (Notebook Toshiba Satellite C660-1NZ+ mikrofon Sennheiser PC26 USB, Notebook Fujitsu Siemens Amilo Pro (dodatkowe nagranie z telefonu), Notebook Compaq 610 + mikrofon Sennheiser PC26 USB)

4. Kanały:

- dźwięk z mikrofonu nagłownego dla obojga rozmówców
- dźwięk zarejestrowany bezpośrednio przez rejestrator w telefonie komórkowym
- obraz z kamery dla obojga rozmówców
- dźwięk z telefonu dla jednego rozmówcy (*inicjujący rozmowę, spk_2*)
 - dodatkowo z telefonu przez laptop nagrywany jest ten rozmówca, który inicjuje rozmowę telefoniczną, on rozmawia „normalnie” przez telefon komórkowy, ten drugi rozmówca mówi przez zestaw zmodyfikowany i jest rejestrowany tylko ze zwykłego mikrofonu, rejestratora w telefonie komórkowym i kamery

5. Pomieszczenia: Collegium Novum, pok. 312ab „BIURO”, pokój 07b „STUDIO”

6. Nazewnictwo katalogów i plików:

- Nagrania każdego mówcy były zapisywane w odrębnym katalogu, nazwanym wg wzoru: rrrrmmdd_nnn, np. 20120604_001 (data_numer). Nazwa katalogu będzie taka sama dla obu mówców danej pary. Mówca w biurze oznaczony jest jako mówca A, mówca w studio - mówca B.
- Nagrania poszczególnych plików nazywane analogicznie wg wzorów:
 - rejestrator **BIURO**:
 - wg wzoru: task n_speakernnn_nnA_headset.wav (nazwa zadania numer_mówca numer mówcy w całym zbiorze_numer sesji rola mówcy (A lub B, w parze) _ rejestrowany kanał), np. task 1_speaker001_01A_headset.wav
 - kamera: task n_speakernnn_nnA_video.avi
 - recmycall: task n_speakernnn_nnA_phone_inc.wav
 - rejestrator **STUDIO**:
 - na laptopie 1: wg wzoru: task n_speakernnn_nnB_headset.wav np. task 1_speaker001_01B_headset.wav
 - na laptopie 2: wg wzoru: task n_speakernnn_nnA_phone_pc.wav np. task 1_speaker001_01A_phone_pc.wav
 - kamera: task n_speakernnn_nnB_video.avi
 - recmycall: task n_speakernnn_nnB_phone_out.wav

7. Instrukcje od osób nagrywających dla rozmówców: oczekiwany jest jak najbardziej naturalny sposób mówienia, więc zasadniczo nie zaleca się

udzielania mówcom wskazówek odnośnie formy wypowiedzi (odnośnie dykcji itp.); przekazuje się jedynie instrukcje techniczne dotyczące zadań do wykonania, starając się tak formułować te instrukcje, aby zachęcić mówców do rozmowy na temat zadania przewidzianego do realizacji, prawdziwym celem jest, aby mówcy byli w miarę możliwości jak najbardziej rozmowni. Osoba nagrywająca może stymulować rozmowę, jeśli zauważy, że rozmowa nie biegnie swobodnie, spontanicznie czy też że mówcom ewidentnie brak pomysłów - zaleca się jednak, aby nagrywający włączał się raczej na zasadzie udziału w rozwiązywaniu zadań, dyskretnych podpowiedzi, aniżeli jawnego nakłaniania, aby uczestnicy po prostu mówili więcej. Kiedy rozmowa toczy się względnie swobodnie - nagrywający nie ingeruje. W nagraniach przeprowadzonych do tej pory wg powyższych instrukcji, istotna ingerencja osób nagrywających w przebieg sesji nie była konieczna (zob. niżej).

3.2.3 Scenariusze nagraniowe

Część 1. Tekst czytany

Mówca czyta tekst „Północny wiatr i słońce” (tempo normalne, nie narzucamy żadnego specjalnego sposobu czytania).

Część 2. „Znajdowanie różnic” (różne obrazki)

Informacja wstępna

Rozmówcy otrzymują po jednym obrazku, na którym przedstawiony jest budynek w przekroju. Rysunki różnią się szczegółami.

Zadanie (Instrukcja, którą przekazują nagrywający)

Otrzymaliście bardzo podobne obrazki, różniące się szczegółami. **RAZEM** z rozmówcą postaraj się znaleźć jak największą liczbę różnic między tymi obrazkami. **WSPÓLNIE** uzgodnijcie moment, gdy uznacie, że to już wszystkie różnice i nic więcej już nie dostrzegacie.

Oczekiwane rezultaty: Elementy dialogu, pytania, odpowiedzi, zawahania, wyrażenie zgody/zaprzeczenie, wyrażenie satysfakcji/niezadowolenia. Tempo mowy normalne. Słownictwo związane z opisem budynku - wewnątrz i na zewnątrz.

Część 3. „Znajdowanie różnic” (takie same obrazki)

Informacja wstępna

W tym zadaniu zależy nam na względnie naturalnym przyspieszeniu tempa mowy. Obaj rozmówcy otrzymują identyczny obrazek, na którym widać dwa zdjęcia pokoju, różniące się szczegółami.

Zadanie (Instrukcja, którą przekazują nagrywający)

Tym razem otrzymaliście taki sam obrazek. Postarajcie się **WSPÓLNIE** ustalić maksymalną liczbę różnic między górnym i dolnym zdjęciem. **UWAGA:** W tym zadaniu liczy się **CZAS**, proszę postarać się jak najszybciej zidentyfikować różnice.

Oczekiwane rezultaty: Elementy dialogu, pytania, odpowiedzi, zawahania, wyrażenie zgody/zaprzeczenie, wyrażenie satysfakcji/niezadowolenia. Tempo mowy normalne/przyspieszone. Słownictwo związane z opisem wnętrza pomieszczenia mieszkalnego.

Część 4. Rysopis osoby cz.1

Informacja wstępna

W tym zadaniu mierzymy czas odpowiedzi. Celem jest podanie jak największej liczby szczegółów w jak najkrótszym czasie. Jeden z rozmówców ma przed sobą tylko jedno zdjęcie, drugi na początku dopytuje o szczegóły i słucha, a następnie otrzymuje do obejrzenia kilka zdjęć różnych postaci w celu identyfikacji osoby wcześniej opisywanej.

Część 4. Rysopis osoby cz.1

Jedno z Was otrzymało zdjęcie osoby. Drugie ma za zadanie uzyskać **JAK NAJWIĘCEJ INFORMACJI** na temat tej osoby, zadając **jak najwięcej PYTAŃ**, dotyczących rysopisu tej osoby, zarówno fizycznych cech, jak i ubioru, stylu uczesania, Państwa skojarzeń, może podobieństw do kogoś ze znanych postaci lub z Państwa znajomych itp. Osoba która widzi zdjęcie może także sama podawać/usupełniać szczegóły rysopisu, o które nie została zapytana - **WSPÓLNIE** mają Państwo ustalić maksymalną liczbę szczegółów wyglądu tej osoby.

- Następnie osoba, która nie widzi obrazka otrzyma pewne **ZADANIE** do wykonania, związane z tym obrazkiem.
- **UWAGA:** W tym zadaniu także liczy się **CZAS**, proszę postarać się jak najszybciej ustalać rysopis, pomagać sobie nawzajem, aby zrobić to jak najsprawniej.

Oczekiwane rezultaty: Elementy dialogu, pytania, odpowiedzi, zawahania, wyrażenie zgody/zaprzeczenie, wyrażenie satysfakcji/niezadowolenia. Tempo mowy normalne/przyspieszone. Słownictwo związane z rysopisem osoby (osób) widocznej na zdjęciu.

Część 5. Rysopis osoby cz. 2

Informacja wstępna

W tym zadaniu również mierzymy czas odpowiedzi. Celem jest podanie jak największej liczby szczegółów w jak najkrótszym czasie. Zadanie podobne do poprzedniego, mówcy zamieniają się rolami. Jeden rozmówca widzi zdjęcie mężczyzny, drugi wypytuje o szczegóły jego rysopisu, razem ustalają ten rysopis. Następnie drugi rozmówca otrzymuje do identyfikacji 3 zdjęcia tego samego mężczyzny, ale w nieco różnych pozach (inny układ rąk, głowy).

Zadanie

Jedno z Was otrzymało zdjęcie osoby. Drugie ma za zadanie uzyskać JAK NAJWIĘCEJ INFORMACJI na temat tej osoby, zadając jak najwięcej PYTAŃ, dotyczących rysopisu tej osoby.

- Następnie osoba, która nie widzi obrazka otrzyma do wykonania ZADANIE związane z tym obrazkiem. Zadanie będzie nieco **INNE** niż w poprzedniej części, więc proszę zachować czujność:)
- **UWAGA:** W tym zadaniu także liczy się **CZAS**, proszę postarać się jak najszybciej ustalać rysopis, pomagać sobie nawzajem, aby zrobić to jak najsprawniej.

Oczekiwane rezultaty: J.w. mowa / elementy dialogu w tempie normalnym/przyspieszonym, słownictwo związane z podawaniem rysopisu osoby (osób) widocznej na zdjęciu, możliwe słownictwo dotyczące mimiki twarzy, gestów.

Część 6. Tekst czytany

Mówca czyta ponownie tekst „Północny wiatr i słońce” (nie narzucamy żadnego specjalnego sposobu czytania).

3.2.4 Przykładowe nazewnictwo plików w sesji nagraniowej

Uzyskano 30 sesji nagranych, nagrania zgromadzono w katalogach, nagrania każdej pary mówców znalazły się w odrębnym katalogu

Nazwa sesji obejmującej nagrania każdej pary rozmówców wg wzoru: rrrmmdd_nnn, np. 20120604_001 (data_kolejny nr sesji). Wewnątrz sesji nazwy plików z objaśnieniem:

- read 1_speaker001_01A_headset.wav - pierwsze czytanie tekstu, mówca 1, para pierwsza, mikr. nagłówny

- **read 1_speaker001_01B_headset.wav** - pierwsze czytanie tekstu, mówca 2, para pierwsza, mikr. nagłowny
- **read 2_speaker001_01A_headset.wav** - drugie czytanie tekstu, mówca 1, para pierwsza, mikr. nagłowny
- **read 2_speaker001_01B_headset.wav** - drugie czytanie tekstu, mówca 2, para pierwsza, mikr. nagłowny
- **task 1_speaker001_01A_headset.wav** - zad. dialogowe 1 (różnice, zad. 1), para pierwsza, mówca 1, mikr. nagłowny

3.2.5 Mówcy

Podstawowe informacje o mówcach zawarto w tabeli poniżej. Mówcy to studenci i pracownicy Wydziału Neofilologii UAM w Poznaniu w wieku między 20 a 35 lat. Większość uczestników nagrań to kobiety. Pogrubioną czcionką oznaczono pary wybrane do dalszej analizy. Pozostałe pary mówców mogą być również przydatne do przeprowadzenia niektórych rodzajów analiz, jednakże nie dla wszystkich, ze względu na jakość nagrania lub inne specyficzne cechy mówców (np. mowa nienatywna, słabsza jakość nagrań niektórych fragmentów lub kanałów).

lp.	ID pary	płeć / wiek mówca 1 (spk1)		płeć / wiek mówca 2 (spk2)	
1.	20120606_001	F	23	M	21
2.	20120606_002	F	22	F	22
3.	20120606_003	F	23	F	20
4.	20120606_004	F	20	F	20
5.	20120606_005	F	20	M	20
6.	20120606_006	M	20	M	20
7.	20120606_007	F	20	F	25
8.	20120606_008	F	20	F	20
9.	20120606_009	F	20	F	21
10.	20120628_001	F	29	F	31
11.	20120628_002	F	29	F	22
12.	20120628_003	F	29	M	28
13.	20120628_004	F	34	M	34
14.	20120628_005	F	25	M	26

15.	20120628_006	M	28	F	35
podsumowanie		13xF, 2xM		9xF, 6xM	

3.3 Mowa nacechowana emocjonalnie - podzbiór „Emo”

Podzbiór “Emo” bazy “Paralingua” składa się z trzech części: nagrań mowy emocjonalnej zrealizowanej przez 9 mówców (w tym 5 profesjonalnych aktorek, 2 aktorów i 2 mówców nieprofesjonalnych), scenki w sądzie oraz imitacji mowy emocjonalnej. Wszystkie nagrania zrealizowano w profesjonalnym studio nagraniowym, pod okiem zawodowego realizatora dźwięku. Nad merytoryczną zgodnością pozyskiwanych nagrań czuwał ekspert - fonetyk.

3.3.1 Nagrania mowy emocjonalnej - portrety emocji

Pierwsza część korpusu została zaprojektowana zgodnie z poniższymi wytycznymi:

1. Materiał językowy powinny stanowić naturalne, a nie „bezsensowne” zdania, co pozwoli na uniknięcie przerysowania ekspresji głosowej emocji i nienaturalnej prozodii, a także ułatwi umiejscowienie wypowiedzi w kontekście emocjonalnym. Ponadto, zdania języka naturalnego łatwiej mogą być zapamiętane, tak więc aktorzy nie będą czytać tekstu z kartek, co pozwoli na bardziej spontaniczną i naturalną realizację wypowiedzi. Zdania powinny być emocjonalnie neutralne, tzn. powinny mieć taką strukturę gramatyczną i zawierać takie słownictwo, które nie będą determinowały przekazu paralingwistycznego wypowiedzi. Powinny być to zdania, o wysokiej częstotliwości użycia w codziennej komunikacji.
2. Materiał językowy powinien reprezentować emocje podstawowe, emocje różniące się pod względem aktywacji, a także emocje pozytywne i negatywne.
3. Liczba mówców nie powinna być zanadto ograniczona, aby można było uogólnić wnioski z analizy akustycznej wypowiedzi emocjonalnych. Wszyscy mówcy powinni zrealizować ten sam materiał językowy we wszystkich badanych emocjach.

Omawiana tutaj część bazy zawiera łącznie 540 wypowiedzi, tj. 5 neutralnych semantycznie i składniowo zdań (patrz poniżej), zrealizowanych przez 9 mówców w 12 stanach emocjonalnych. Przed rozpoczęciem nagrań, aktorzy otrzymali materiały zawierające definicję każdej z 12 emocji oraz 2 scenariusze dla każdej emocji, które miały ułatwić wzbudzenie emocji w mówcach i zastosowanie odpowiednich technik gry aktorskiej.

Teraz wszystko rozumiem.
Dzisiaj jest poniedziałek.
Od rana pada deszcz.
Powiedział, że nic się nie stało.
Jedziemy na wycieczkę do Grecji.

3.3.2 Imitacja rozpraw sądowych

Druga część korpusu "Emo" to nagrania scenki w sądzie inspirowane popularnymi obecnie serialami paradokumentalnymi. Ta część bazy zawiera mowę afektywną spontaniczną, cechującą się znacznie większą naturalnością niż nagrania zdań. Aktorzy otrzymali skrypt, zawierający dokładny opis sytuacji, osób i sprawy o rozwód, która miała się toczyć w sądzie. Skrypt został napisany w formie didaskaliów - aktorzy na ich podstawie wiedzieli co mają powiedzieć, ale sami decydowali o strukturze językowej i pozajęzykowej wypowiedzi. W nagraniach wzięły udział dwie aktorki (sędzina i powódka, żona) i aktor (pozwany, mąż). Nagrano dwie scenki, jedną z udziałem „żony”, drugą z udziałem „męża”, składających zeznania przed sędzią. W sumie uzyskano 20 minut nagrań mowy spontanicznej, interaktywnej (z uwagi na dialogi), a także pozwalającej na analizę ekspresji emocji i innych stanów afektywnych o różnym stopniu natężenia i w szerszym kontekście czasowym.

3.3.3 Naśladowane emocje rzeczywiste

Trzecia część korpusu zawiera imitacje spontanicznych wypowiedzi emocjonalnych zrealizowanych przez 9 mówców (w tym 5 profesjonalnych aktorek, 2 aktorów i 2 mówców nieprofesjonalnych). Oryginalne nagrania pochodzą z policyjnej bazy 997, z której wybrano 20 zgłoszeń, dokonanych dwukrotnie przez każdego z 10 mówców (5 kobiet i 5 mężczyzn), przy czym jedno z nagrań było wyraźne nacechowane negatywnym stanem emocjonalnym, zaś drugie było neutralne, lub znacznie odmienne pod względem intensywności emocji od pierwszego.

Aktorzy najpierw wysłuchiwali wypowiedzi autentycznych w całości, a następnie realizowali je fragment po fragmencie w sposób jak najbardziej zbliżony do oryginału (pod względem prozodii, słownictwa i emocji).

3.4 Analiza zawartości bazy

3.4.1 Uwagi ogólne

Podzbiór „Dial” - dialogi zadaniowe

Na podstawie wstępnego odsłuchania wyżej opisanych nagrań przez zespół ekspertów Fundacji UAM, stwierdzono, że nagrania cechują się stosunkowo wysokim

stopniem spontaniczności, większość rozmówców sprawia wrażenie rozluźnionych i swobodnie reagujących na wyznaczone zadania.

W części dialogowej, mówcy zgodnie z oczekiwaniami koncentrują się raczej na rozwiązywaniu przez siebie zadaniu, aniżeli na samym fakcie bycia nagrywanym, w znakomitej większości przypadków bardzo dobrze współpracują ze sobą w celu realizacji zadania, pomagając sobie nawzajem. Wykazują przy tym indywidualne zróżnicowanie w sposobie prowadzenia dialogu (np. różnice dominacja vs. ustępowanie/spolegliwość, częstość i sposób przejmowania kolejki dialogowej, narzucenie bądź nie sposobu rozwiązania zadania czy analizy poszczególnych przypadków).

W przypadku części par rozmówców już wstępna analiza pozwala stwierdzić, iż udało się uzyskać efekt przyspieszonego tempa mowy w częściach 3, 4, 5 sesji nagraniowej (zob. wyżej, rozdział pt. Scenariusze nagraniowe), w pozostałych przypadkach wymagana jest dokładniejsza analiza tego efektu.

Ponadto, zebrany materiał oceniono jako przydatny do analizy cech suprasegmentalnych dłuższych wypowiedzi o charakterze spontanicznym i quasi-spontanicznym, specyficznych cech rytmicznych wypowiedzi w języku polskim, jak również pewnych wybranych cech leksykalnych i składniowych współczesnej polszczyzny mówionej. W związku z tym, że mówcy realizują w sposób indywidualny analogiczne zadania, możliwym będzie uzyskanie informacji na temat specyficznych strategii stosowanych przez poszczególnych mówców dla osiągnięcia wyznaczonego celu.

Dotychczas prowadzone prace koncentrują się przede wszystkim na analizie materiału dźwiękowego, nagrania z kamer video traktowane są jako materiał pomocniczy, niemniej jednak w dalszej perspektywie planuje się również pilotażową analizę materiałów wizualnych na potrzeby badania korelacji mimiki twarzy i gestów oraz cech paralingwistycznych komunikatu słownego.

Podzbiór „Emo”

Zarejestrowano nagrania wysokiej jakości, w profesjonalnym studio nagraniowym, dzięki czemu możliwe są zarówno percepcyjne analizy materiału, jak również wszelkie analizy instrumentalne.

Na podstawie analiz przeprowadzonych przez ekspertów, jak również w ramach pilotażowego eksperymentu z udziałem słuchaczy naiwnych, można potwierdzić przydatność zgromadzonych danych w procesie opisu i modelowania emocji w głosie.

Dzięki reprezentacji stosunkowo szerokiego wahlarza emocji, a także wykorzystaniu opisanych wyżej technik pozyskania materiału dźwiękowego (mowa nie jest bezpośrednio czytana, odgrywano emocje na podstawie zarysów sytuacji, a nie bezpośrednich definicji emocji) udało się uzyskać równowagę pomiędzy względną naturalnością i autentyzmem emocji wypowiedzi oraz kontrolą i porównywalnością nagrań. Badanie tej równowagi możliwe jest również dzięki

zarejestrowanym wypowiedziom aktorskim naśladowującym rzeczywiste emocje z wypowiedzi z telefonu alarmowego.

Jako punkt odniesienia odnośnie cech głosu przy realizacji wypowiedzi nacechowanych emocjonalnie oraz neutralnych mogą też służyć dodatkowe nagrania w podzbiorze „Emo”, obejmujące teksty czytane w tempie wolnym oraz szybkim, a także imitowana rozprawa sądowa (elementy dialogu oraz pokrycie wybranych struktur leksykalnych i składniowych).

3.4.2 Transkrypcja i anotacja

Przeprowadzono transkrypcję ortograficzną i anotację wszystkich zarejestrowanych nagrań z wykorzystaniem programu Annotation System (informacje dotyczące narzędzia, zob. niżej). Dla każdej nagranej wypowiedzi utworzono plik o rozszerzeniu .ant (z anotacją w formacie XML).

W przypadku nagrań tekstów czytanych transkrypcja została przeprowadzona pół-automatycznie: importowano tekst wraz z wstępnym podziałem na frazy z pliku w formacie .txt, a następnie manualnie dokonano korekty tekstu wypowiedzi (identyfikacja i oznaczenie odstępstw od oryginalnego tekstu) oraz korekty segmentacji nagrań. Na tym etapie podział na frazy został przeprowadzony na podstawie subiektywnej oceny osób anotujących (eksperti z dziedziny fonetyki), za wyznaczniki granic fraz uznawano przede wszystkim pauzy oraz w drugiej kolejności wskazówki prozodyczne.

Nagrania mowy spontanicznej transkrybowano i anotowano manualnie na poziomie ortograficznym. Dokonywano bieżącej analizy błędów wypowiedzi, pomyłek językowych, dodatkowych dźwięków wydawanych przez mówcę, pauz wypełnionych oraz zakłóceń zewnętrznych i oznaczano je na odrębnych warstwach anotacji.

Anotacja prozodii oraz cech parajęzykowych

Na potrzeby anotacji i analizy cech prozodycznych, parajęzykowych oraz niejęzykowych z wykorzystaniem specyfikacji opracowanych przez zespół Fundacji UAM zaprojektowano i stworzono nowe narzędzie: *Annotation System*, opisane w rozdziale 4 poniżej. Umożliwia ono zarówno anotację z wykorzystaniem tradycyjnych warstw anotacji oraz systemów etykiet (np. wg SPEECON czy BLF), jak również anotację cech ciągłych z zastosowaniem graficznej reprezentacji przestrzeni cech (zob. w szczególności p. 4.6.3, a także publikację Klessa, Wagner, Oleśkiewicz-Popiel, 2013).

4 Annotation System. Oprogramowanie do anotacji nagrań mowy spontanicznej

4.1 Informacje ogólne

Program Annotation System umożliwia odtwarzanie dźwięku, wyświetlanie obrazu dźwięku w postaci wykresu amplitudowo-czasowego oraz spektrogramu, a także anotację nagrań. Program w wersji podstawowej przeznaczony jest do manualnej transkrypcji i anotacji z użyciem dowolnej liczby warstw anotacji, jednak jego funkcjonalność może być praktycznie dowolnie rozszerzana za pomocą wtyczek (pluginów) napisanych w języku C#, możliwa jest więc zatem automatyzacja procesu anotacji poprzez takie wtyczki. Ze względu na elastyczną architekturę oraz konfigurowalny interfejs użytkownika w programie możliwa jest realizacja bardzo różnych, indywidualnie wyspecjalizowanych specyfikacji anotacji.

4.2 Architektura programu

Annotation System został napisany w nowoczesnym języku C# w środowisku Visual Studio. Struktura systemu została opracowana jako architektura wielowarstwowa, gdzie każda warstwa aplikacji reprezentuje pewną wirtualną warstwę funkcjonalną. W systemie występują następujące warstwy:

Rysunek 1: Warstwy oprogramowania Annotation System

Rozdzielenie warstw ze względu na ich funkcjonalność zapewnia przyszły rozwój aplikacji i dostosowanie jej w przyszłości do różnych systemów operacyjnych oraz dowolnych baz danych. Annotation System został napisany w pełni obiektowo i został zoptymalizowany do czytania dużych plików.

W systemie od podstaw została zaimplementowana obsługa plików audio (wav), dzięki czemu można było w dowolny sposób modyfikować obsługę i odtwarzanie plików dźwiękowych z dokładnością co do milisekund.

Warstwa Database reprezentuje elementy zapewniające zapis i odczyt danych na poziomie najniższym. W tej chwili jest to zapis i odczyt anotacji do formatu XML oraz obsługa własnego formatu ANT, który jest w rzeczywistości archiwum ZIP i zawiera wewnątrz spakowany plik XML z anotacją. Takie rozwiązanie zapewnia nam w przyszłości możliwość osadzania dowolnej treści wewnątrz pliku ANT. Otwieraniem i zapisywaniem tego formatu zajmuje się w tej chwili warstwa Database. W przyszłości istnieje możliwość modyfikacji tej warstwy, aby można było czytać dane do programu bezpośrednio z bazy danych (SQL Server, ORACLE lub MySQL).

Warstwa Logic jest warstwą pośrednią reprezentującą dane w postaci obiektów w języku C#, którymi możemy się posługiwać z poziomu programisty w celu dokonywania operacji na obiektach oraz kolekcjach obiektów w aplikacji: Layer, Segment, Speaker.

Warstwa Presentation jest warstwą najwyższego poziomu i zawiera kontrolki, które można bezpośrednio umieścić na formularzu i które reprezentują elementy interfejsu: Spectrogram, Waveform, Layer Collection, Input Device. Wszystkie te elementy potrafią współpracować ze sobą automatycznie umożliwiając programiście w kilka minut stworzenia dowolnego klonu aplikacji w oparciu o funkcjonalność Annotation System. Kontrolki warstwy Presentation pełnią funkcję komponentów, które są ze sobą połączone poprzez specjalną kontrolkę - Synchronizer. Synchronizer jest specjalnym obiektem, który przechowuje stan zmiennych, których synchronizacja jest niezbędna w celu spójnego działania programu. Umieszczenie tych elementów w jednym obiekcie porządkuje projekt i sprawia, że staje on się odporny na błędy programisty.

Plugin - jest to fragment zawierający obsługę wtyczek. Każda osoba znająca język programowania C# może nawet w kilka minut napisać dowolną wtyczkę rozszerzającą funkcjonalność Annotation System. Program podczas uruchomienia skanuje folder Annotation System Plugins, znajdujący się w Dokumentach Użytkownika oraz tworzy w menu Plugins listę wtyczek, które znalazł w folderze. Wtyczki są zwykłymi plikami C# (*.cs), które powinny mieć odpowiednią strukturę, załączoną w przykładowej wtyczce, dostępnej dla użytkownika, również w folderze Annotation System Plugins. Programista ma swobodny dostęp do wszystkich kontrolek zastosowanych w interfejsie, do warstw anotacji oraz do danych. Użytkownik w kilka chwil może zmodyfikować aplikację tworząc wtyczkę i potrzebnej funkcjonalności. Może również rozszerzyć system o swoje własne funkcje. W przyszłości warto rozpatrzyć utworzenie serwisu internetowego, gdzie użytkownicy z poziomu systemu będą mogli umieszczać swoje wtyczki oraz korzystać z wtyczek innych użytkowników.

Shared - jest to biblioteka zawierająca drobne klasy pomocnicze wymagane przez wiele warstw.

Sposób połączenia kontrolek i danych Annotation System przedstawiono na rycinie poniżej.

Rysunek 2: Połączenie kontrolek i danych w Annotation System

4.3 Instalacja i uruchomienie programu

Do działania programu niezbędne są następujące elementy:

- Windows Installer 3.1
- Microsoft .NET Framework 4 (x86 and x64)

4.4 Rozpoczęcie pracy

Opcje dotyczące plików z dźwiękiem znajdują się w górnej części okna programu (otwieranie pliku audio, odtwarzanie, zatrzymanie odtwarzania, zapętlenie (loop), zoom itp.), natomiast opcje odnoszące się specyficznie do operacji związanych z anotacją (otwarcie istniejącego pliku anotacji, dodawanie/usuwanie warstw i segmentów) - w dolnej części okna. W prawej górnej części okna programu znajduje się kontrolka graficzna umożliwiająca anotację cech ciągłych z wykorzystaniem reprezentacji graficznej płaszczyzny cech.



FUNDACJA UNIWERSYTETU
IM. ADAMA MICKIEWICZA
W POZNANIU

ul. Rubież 46, 61-612 Poznań **tel.** 061 827 97 00, fax 061 827 97 01
e-mail: fundacjauam@ppnt.poznan.pl www.fundacja.ppnt.poznan.pl
KRS: 0000052045 **Regon:** 631509229 **NIP:** 781-00-02-075



POZNAŃSKI PARK
NAUKOWO-TECHNOLOGICZNY

Fundacji Uniwersytetu im. A. Mickiewicza



FUNDACJA UNIWERSYTETU
IM. ADAMA MICKIEWICZA
W POZNANIU

ul. Rubież 46, 61-612 Poznań **tel.** 061 827 97 00, fax 061 827 97 01
e-mail: fundacjauam@ppnt.poznan.pl www.fundacja.ppnt.poznan.pl
KRS: 0000052045 **Regon:** 631509229 **NIP:** 781-00-02-075



POZNAŃSKI PARK
NAUKOWO-TECHNOLOGICZNY

Fundacji Uniwersytetu im. A. Mickiewicza



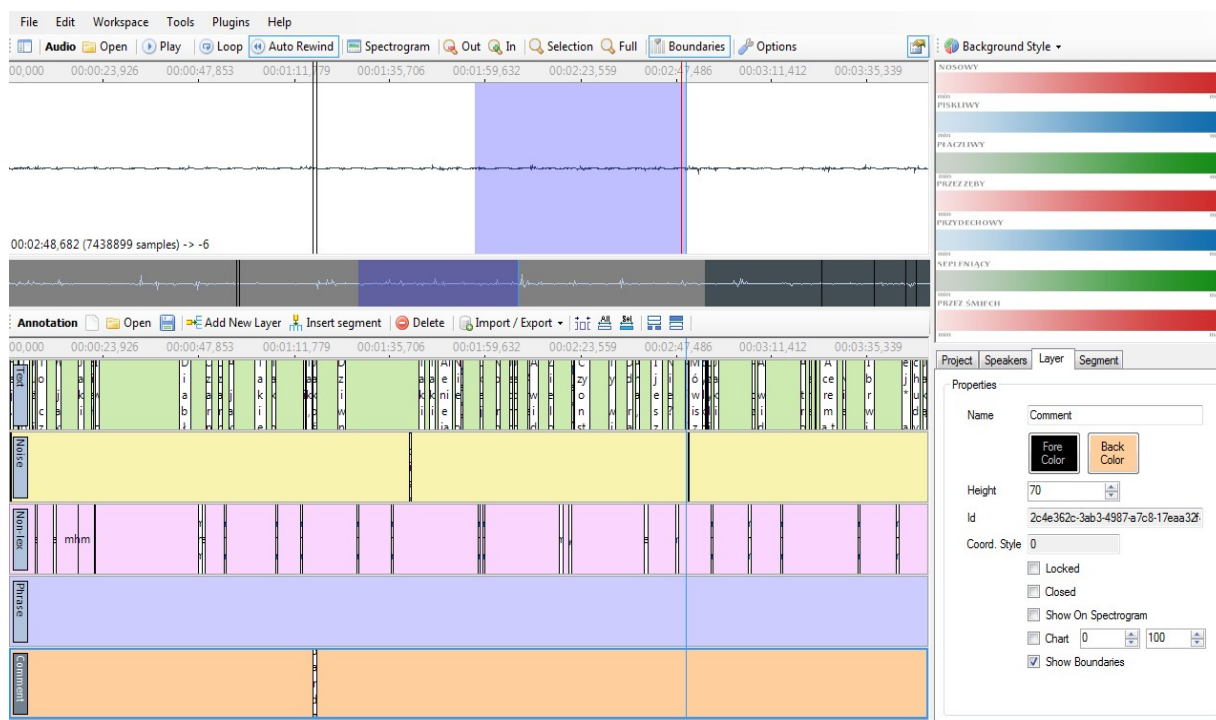
FUNDACJA UNIWERSYTETU
IM. ADAMA MICKIEWICZA
W POZNANIU

ul. Rubież 46, 61-612 Poznań **tel.** 061 827 97 00, fax 061 827 97 01
e-mail: fundacjauam@ppnt.poznan.pl www.fundacja.ppnt.poznan.pl
KRS: 0000052045 **Regon:** 631509229 **NIP:** 781-00-02-075



POZNAŃSKI PARK
NAUKOWO-TECHNOLOGICZNY

Fundacji Uniwersytetu im. A. Mickiewicza



Ilustracja 2: Okno programu Annotation System

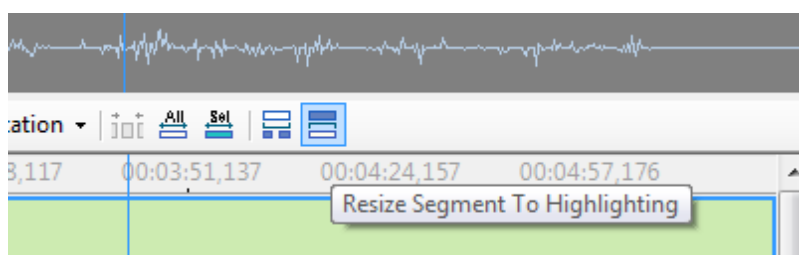
Przy pracy na plikach znajdujących się w katalogu na dysku, a pliki z nagraniem oraz z anotacją (.ant) mają takie same nazwy (różnią się wyłącznie rozszerzeniami), to po naciśnięciu przycisku „Open” i wskazaniu pliku z anotacją, program zapyta czy ma otworzyć również plik wave o takiej samej nazwie.

4.5 Operacje na segmentach i warstwach anotacji

Segmnty na warstwie anotacji i operacje na tekście w trakcie segmentacji:

- aby utworzyć segment należy kliknąć na warstwie, na której segment ma być wstawiony i zaznaczamy myszą ten fragment nagrania, dla którego chcemy utworzyć nowy segment, a następnie naciskamy klawisz *Insert* na klawiaturze (lub przycisk *Insert segment*)
- aby wpisać treść tekstu do segmentu należy rozpocząć pisanie po wstawieniu segmentu, tekst wyświetla się od razu
- w celu modyfikacji tekstu już istniejącego segmentu (po wyjściu z trybu edycji), możemy dwukliknąć segment lub użyć pola tekstowego *label* na zakładce *Segment* po prawej stronie okna programu:

- zaznaczenie segmentu: należy dwukliknąć segment, co powoduje zaznaczenie odpowiadającego mu fragmentu nagrania
- odtwarzanie dźwięku dla wybranego segmentu: **CTRL spacja** (aby zatrzymać - ponownie ta sama kombinacja), dzięki temu możliwe jest odtwarzanie dźwięku także w trybie pisania tekstu
- zmiana szerokości segmentu:
 - należy nadjechać kursorem myszki nad jego granicę, klikamy i rozciągamy segment
 - zmiana rozmiaru poprzez wyrównanie do zaznaczenia na wave: należy a) zaznaczyć na wave interesujący nas fragment nagrania, b) kliknąć na segment, którego rozmiar chcemy wyrównać do zaznaczenia, c) kliknąć przycisk *Resize Segment to Highlighting*



Ilustracja 3: Zmiana rozmiaru segmentu zgodnie z zaznaczeniem

- zmiana położenia segmentu na warstwie i przestawienie go na inną warstwę: po nadjechaniu nad segment (NIE nad krawędź), można go „złapać myszką” i przesuwać w dowolne miejsce
- wyświetlenie wszystkich wstawionych etykiet anotacji na jednym ekranie: Zoom to All Segments:



Ilustracja 4: Wyświetlenie wszystkich etykiet anotacji na jednym ekranie.

- zoom (zbliżenie) - najwygodniejszy jest „scroll” myszki: po kliknięciu na przebiegu czasowym należy użyć „scroll”: powoduje to zbliżanie / oddalanie się do/od miejsca w którym stoi kursor; poza tym dostępne są przyciski zmieniające zbliżenie „In” / „Out”
- przesuwanie całego przebiegu i warstw anotacji: możemy ciągnąć za linijkę nad anotacją, kursor myszy zamienia się nad tą linijką w „łapkę”
- drugim nawigatorem jest podgląd przebiegu czasowego - na nim widzimy: obecne zaznaczenie oraz obecny zoom, za który możemy te ciągnąć i przesuwać widok naszego nagrania i anotacji:

(zob niżej.: niebieskie zaznaczenie to zaznaczenie na przebiegu czasowym, szare zaznaczenie pokazuje fragment widoczny na ekranie, za szare zaznaczenie możemy ciągnąć, jego szerokość zależy od zbliżenia (zoom'u))



Ilustracja 5: Pasek nawigacji (przesuwania zawartości) i regulacji zbliżenia (zoom)

- blokada i minimalizacja warstw:
 - **blokada warstwy:** każdą warstwę anotacji można zablokować (edycja jest wówczas niemożliwa, ale segmenty są normalnie widoczne), ma to zapobiec przypadkowym modyfikacjom danej warstwy; w celu zablokowania warstwy należy zaznaczyć opcję „Locked” na zakładce „Layer” po prawej stronie okna programu:



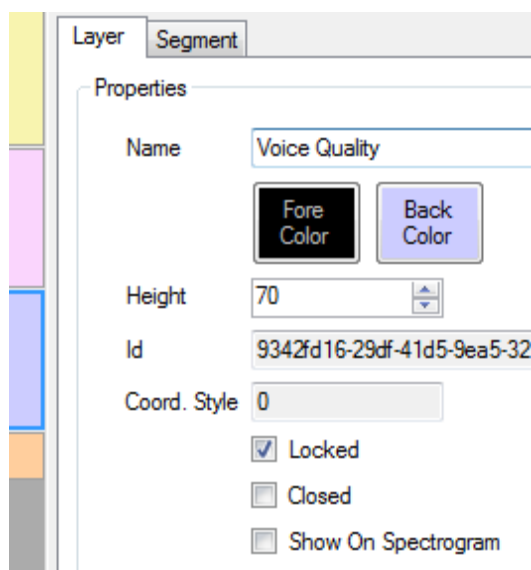
FUNDACJA UNIWERSYTETU
IM. ADAMA MICKIEWICZA
W POZNANIU

ul. Rubież 46, 61-612 Poznań **tel.** 061 827 97 00, fax 061 827 97 01
e-mail: fundacjauam@ppnt.poznan.pl www.fundacja.ppnt.poznan.pl
KRS: 0000052045 **Regon:** 631509229 **NIP:** 781-00-02-075



POZNAŃSKI PARK
NAUKOWO-TECHNOLOGICZNY

Fundacji Uniwersytetu im. A. Mickiewicza



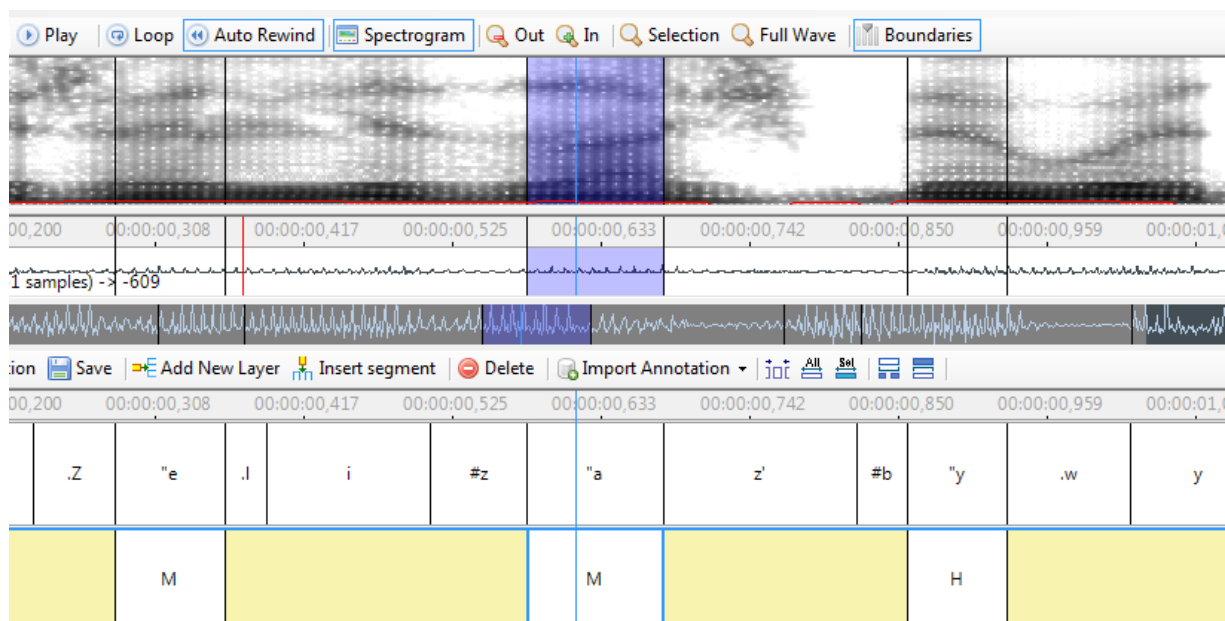
Ilustracja 6: Blokada edycji na warstwie anotacji

- o **blokada i minimalizacja warstwy:** każdą warstwę można także zminimalizować (jednocześnie jest ona wówczas także zablokowana i nie można jej edytować), na ekranie wyświetla się wówczas jedynie nazwa warstwy i zwężony pasek tej warstwy, co daje więcej możliwości odnośnie wyświetlania danych na ekranie, ponieważ warstwa zajmuje mniej miejsca; warstwę można zablokować i zminimalizować dwuklikając na jej nazwie po lewej stronie lub zaznaczając opcję „Closed” na zakładce „Layer” po prawej stronie okna programu.

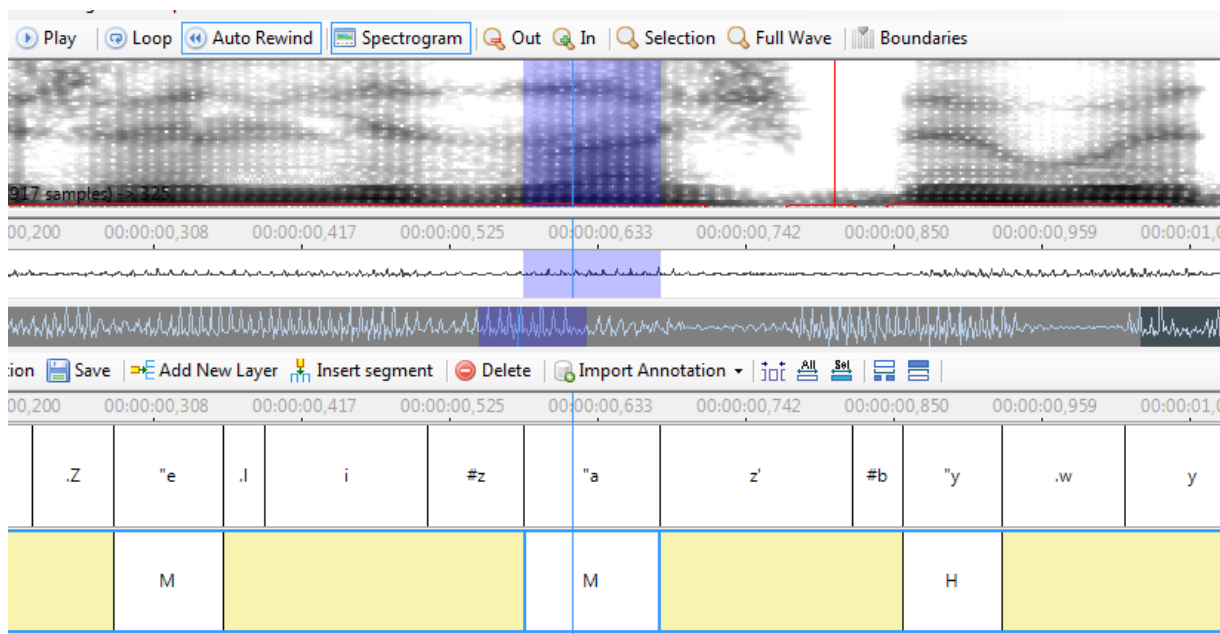


Ilustracja 7: Minimalizacja i blokada warstw anotacji.

- wyświetlanie granic segmentów na obrazie dźwięku - granice segmentów domyślnie są przedłużone, tak aby wyświetlały się także na spektrogramie i wykresie amplitudowo-czasowym; domyślnie wyświetlane są granice z tej warstwy, która aktualnie jest zaznaczona
- ukrywanie granic segmentów - w razie potrzeby można wyłączać / wyłączać wyświetlanie granic na spektrogramie i wykresie czasowym, klikając przycisk „Boundaries”:



Ilustracja 8: Przedłużenie granic segmentów warstwy anotacji na obraz dźwięku - włączone



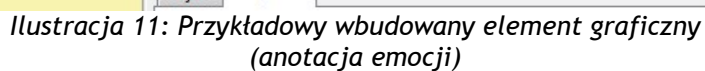
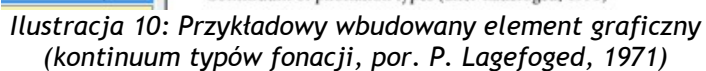
Ilustracja 9: Przedłużenie granic segmentów warstwy anotacji na obraz dźwięku - wyłączone

- opcja ukrycia granic może być szczególnie przydatna w przypadku warstw z bardzo dużą liczbą granic, np. dla warstwy zawierającej chwilowe wartości F0.

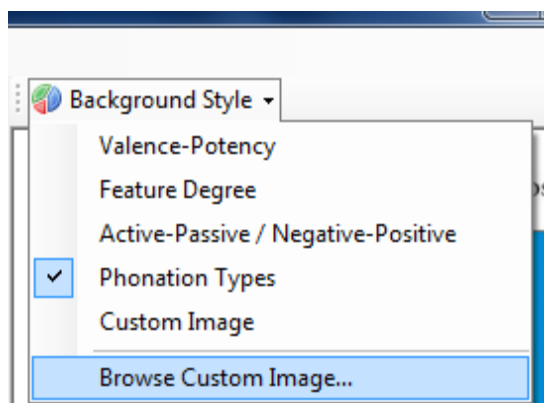
4.5.1 Element graficzny - wspomaganie anotacji cech ciągłych / skalarnych

W prawywej górnej części okna programu znajduje się element graficzny. W program wbudowanych zostało kilka przykładowych elementów, bazujących na wcześniejszych lub obecnych badaniach dotyczących opisu cech mowy za pomocą reprezentacji graficznych (zob. też np. publikacje: Klessa, et al, 2013).

- Elementy te dają możliwość anotacji cech mowy, które trudno jest skategoryzować czy przypisać im konkretne wartości czy poziomy / natężenia.
- W przypadku niekorzystania z opcji elementu graficznego możliwe jest jego ukrycie (należy w tym celu kliknąć przycisk „Properties”)
- Program zapisuje w pliku z anotacją informację o tym, który element graficzny został wykorzystany do anotacji na danej warstwie anotacji (inaczej mówiąc - warstwa „pamięta”, który obrazek wykorzystano do wstawiania współrzędnych na tej warstwie)
- Sposób wykorzystania elementu graficznego jest uzależniony od potrzeb / inwencji / interpretacji użytkownika końcowego, który może wykorzystać wbudowane lub własne elementy graficzne w sposób odpowiadający własnym potrzebom



- w miejscu domyślnego, wbudowanego elementu użytkownik może wstawić dowolną grafikę (jpg, bmp) z zachowaniem rozmiaru, która będzie stanowiła płaszczyznę analizy



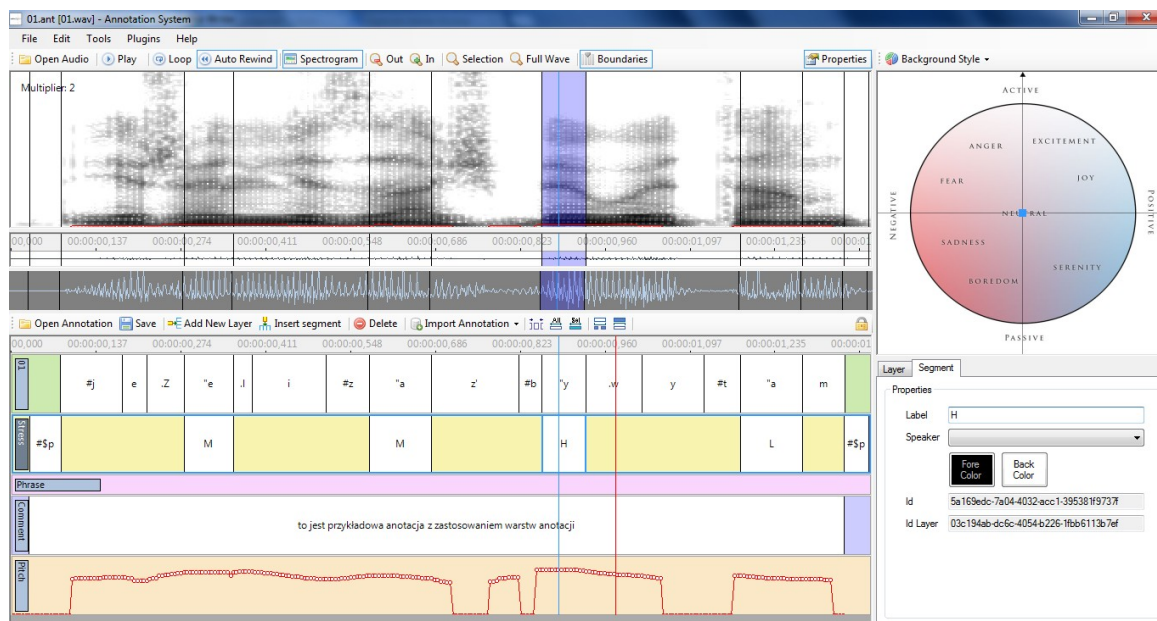
Ilustracja 12: Wybór elementu graficznego

- po kliknięciu na graficznej reprezentacji cech, na zaznaczonej warstwie anotacji wstawiane są współrzędne punktu, w który kliknęliśmy
- po kliknięciu odtwarzania dźwięku i następnie kolejnym kliknięciu dowolnej liczby punktów na obrazku (reprezentacji graficznej) w momentach kliknięcia wstawiane są kolejne segmenty o kolejnych współrzędnych (możliwe jest klikanie i rejestracja położenia punktów, które kliknięto w trakcie odtwarzania)
- kliknięcie z jednoczesnym przytrzymaniem klawisza CTRL powoduje zapisanie kilku położenia punktu w ramach jednego segmentu (np. gdy w opinii anotującego w danym momencie w nagraniu występuje kilka różnych wartości dla danej cechy/cech)

4.5.2 Wyświetlanie wykresów na podstawie zawartości warstwy anotacji

Możliwe jest wykreślenie wykresów na podstawie wartości liczbowych znajdujących się na warstwie anotacji. Wykresy mogą być wyświetlane albo jako krzywe nałożone na spektrogram, albo na odrębnej warstwie anotacji.

- w celu wyświetlenia zawartości warstwy w postaci wykresu nałożonego na spektrogram należy zaznaczyć warstwę z wybraną zawartością (np.

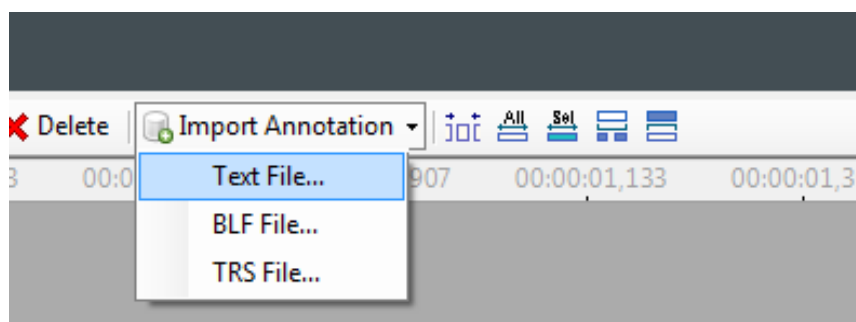


Ilustracja 14: Okno programu. Dolna warstwa przedstawia wykres F0 uzyskany za pomocą zewnętrznej wtyczki (pluginu)

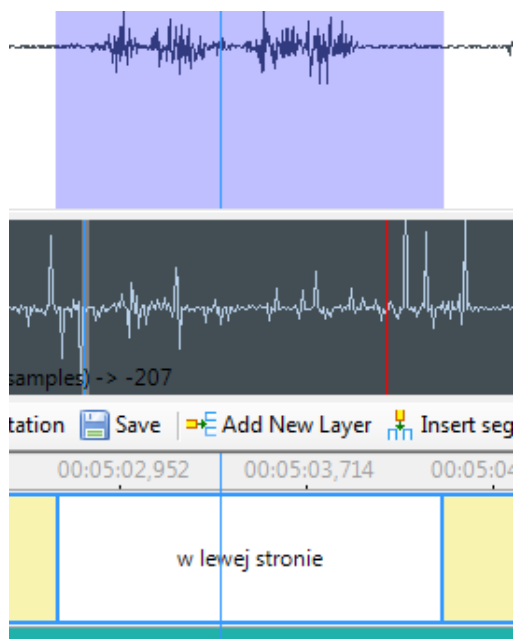
4.6 Przykładowe specyfikacje anotacji w programie i przeprowadzone analizy

4.6.1 Teksty czytane w bazie „Paralingua” - zapis ortograficzny tekstu czytanego, z podziałem na frazy

1. Należy otworzyć plik z nagraniem (przycisk: *Open Audio*) i odsłuchać czytany tekst.
2. W pliku tekstowym zawierającym zapis ortograficzny nagrania (w notatniku) należy podzielić tekst na wiersze (wstawiając znak końca linii) zgodnie z podziałem na frazy zastosowanym przez czytającego (przy podziale na frazy uwzględnia się przede wszystkim występowanie i długość pauz i/lub wyraźną i jednoznaczną zmianą intonacji, wskazującą na silną granicę frazy).
3. W programie *Annotation System* importujemy tekst z pliku tekstowego warstwę anotacji (opcja: *Import Annotation -> Text File*)

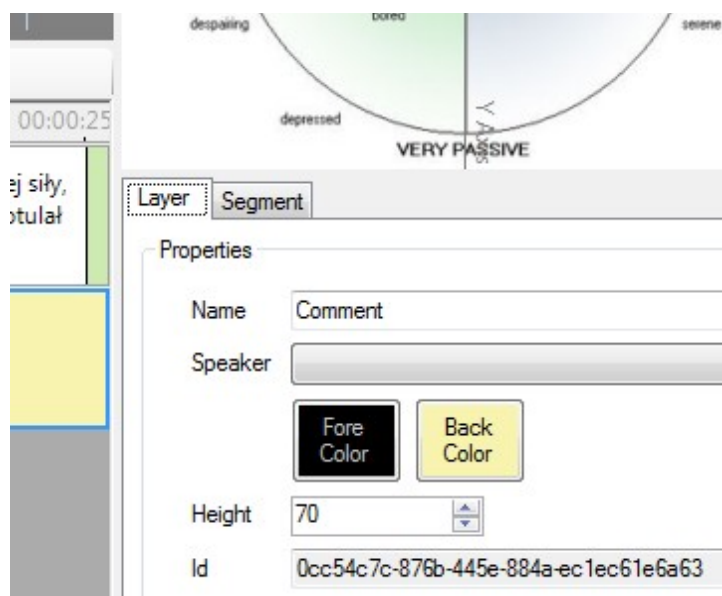


Ilustracja 15: Import anotacji z plików o innych formatach (tu: format pliku TXT, BLF -BOSS label file, TRS - format programu Transcriber)



*Ilustracja 16: Wstawienie
(synchronizacja rozmiaru) segmentu
anotacji z zaznaczeniem*

4. Po zaimportowaniu transkrypcji z pliku tekstowego (program automatycznie sam utworzy nową warstwę anotacji przy imporcie) należy zweryfikować i wyrównać segmenty odpowiadające frazom wydzielonym w tekście, tak aby zgadzały się w czasie z dźwiękiem.
5. W razie potrzeby skomentowania czytanej wypowiedzi można w pliku z anotacją dodać drugą warstwę (przycisk *Add New Layer*) i nadać jej nazwę *Comments* (w panelu po prawej stronie, na pierwszej zakładce, zob. niżej). Na tej warstwie można dodawać swobodne komentarze odnośnie fragmentów czytanych wypowiedzi (np. gdy mówca zmieni sposób czytania, przyspieszy, zwolni, wtrąci coś, kaszlnie, itd. a także w przypadku dźwięków z zewnątrz czy odnośnie jakości nagrania - np. mówca odsunął mikrofon i wyraźnie słyszymy, że poziom głośności dźwięku zdecydowanie odbiega od reszty sesji tego mówcy.).



Ilustracja 17: Boczny panel - opcje dotyczące warstwy

4.6.2 Nagrania dialogowe „Paralingua” - zapis ortograficzny tekstu z podziałem na frazy z wyodrębnieniem innych dźwięków mówcy i otoczenia

1. Otwieramy plik z nagraniem (Przycisk *Open Audio*)
2. Otwieramy plik 000.ant (Przycisk *Open Annotation*)
3. Wybieramy *Zapisz jako...*, a następnie zapisujemy plik anotacji w katalogu razem z nagraniem pod nazwą taką samą, jak nazwa nagrania (różnica tylko w rozszerzeniu - pozostawiamy .ant). W ten sposób przy kolejnym otwieraniu tego samego pliku .ant otworzy się on automatycznie razem z nagraniem.
4. Otwarty plik .ant ma już wstępnie utworzonych 5 warstw:

	00,000	00:00:33,019	00:01:06,039	00:01:39,059	00:02:12,078	00:02:45,
Text						
Noise						
Non-lex						
Phrase						
Comment						

Ilustracja 18: Przykładowy zestaw warstw

Objaśnienie nazw warstw:

- *Text* - tekst wypowiedzi, wszystkie wypowiadane zdania, frazy, wyrazy
- *Noise* - zakłócenia zewnętrzne oraz dźwięki mówcy raczej nie związane bezpośrednio z komunikacyjnym aspektem wypowiedzi (śmiech, naturalny kaszel, oddech, mlaśnięcia itp.).
- *Non-lex* - dźwięki mówcy, które nie są typowymi wyrazami języka polskiego, ale można założyć, że zostały użyte w celu związanym z komunikacją (pauzy wypełnione, mhm, aa, yy, eee, i.in.)
- *Phrase* (w pierwszym etapie nie używamy tej warstwy)
- *Comment* - dowolny komentarz anotującego na temat wskazanego poprzez wstawiony segment fragmentu wypowiedzi (podobnie jak w przypadku opcjonalnej warstwy komentarzy dla tekstów czytanych - wszelkiego rodzaju komentarze)

5. Odsłuch nagrania i zapis słyszanego tekstu:

- W nagraniach często słyszalny jest głos głównego rozmówcy oraz w tle - głos drugiego rozmówcy, treści jego wypowiedzi nie transkrybujemy. Na warstwie *Text* zapisujemy tylko wypowiedzi głównego rozmówcy

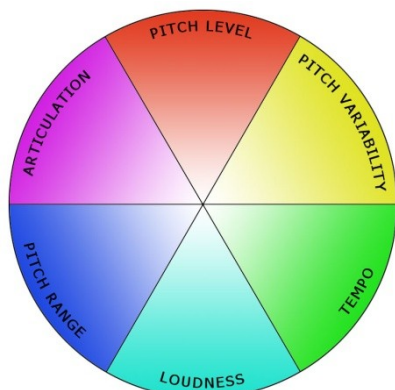
(anotacja dla każdej osoby będzie w odrębnym pliku, nazwanym tak jak nazwa odpowiadającego jej pliku wave). Jeśli głosy mówców nakładają się i głos rozmówcy jest wyraźnie słyszalny w trakcie, gdy mówi główny rozmówca, to zaznaczamy fragment, gdzie mowa się nakłada i wstawiamy segment na warstwie *Noise* lub wpisujemy komentarz na warstwie *Comments*.

- Priorytetowe jest opisanie zdarzeń dotyczących bezpośrednio głównego rozmówcy: **tekstu**, pauz wypełnionych, ewentualnie innych dźwięków przez niego wydawanych. Pozostałe zjawiska słyszalne w nagraniu traktujemy na tym etapie drugorzędnie, jeśli jest możliwość ich oznaczenia na pozostałych istniejących warstwach, należy to zrobić.
- Istnieje możliwość dodania dowolnego komentarza na warstwie *Comments*.

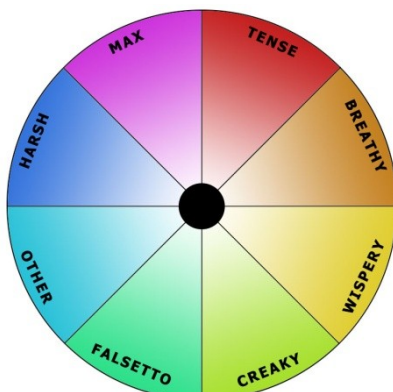
4.6.3 Anotacja z wykorzystaniem graficznej reprezentacji przestrzeni cech

Fonetyczny opis jakości głosu i percepcja prozodii

Przeprowadzono anotację percypowanego przebiegu F0 i prozodii z wykorzystaniem graficznej reprezentacji przestrzeni cech przedstawionych niżej (Ilustracja 19). Na rycinie przedstawionych jest sześć obszarów odpowiadającym cechom: *poziom intonacji*, *zmienność intonacji*, *tempo*, *głośność*, *zakres F0* oraz *artykulacja*. Pomiedzy środkiem koła i jego obwodem widoczny jest dodatkowy mniejszy okrąg odpowiadający nienacechowanej wartości danej cechy, np. normalnemu tempu mowy czy umiarkowanemu zakresowi zmian F0. Środek koła oznaczał wartość minimalną, obwód - maksymalną wartość natężenia danej cechy.



Ilustracja 19: Graficzna reprezentacja przestrzeni cech zastosowana w anotacji prozodii



Ilustracja 20: Graficzna reprezentacja przestrzeni cech zastosowana w anotacji jakości głosu

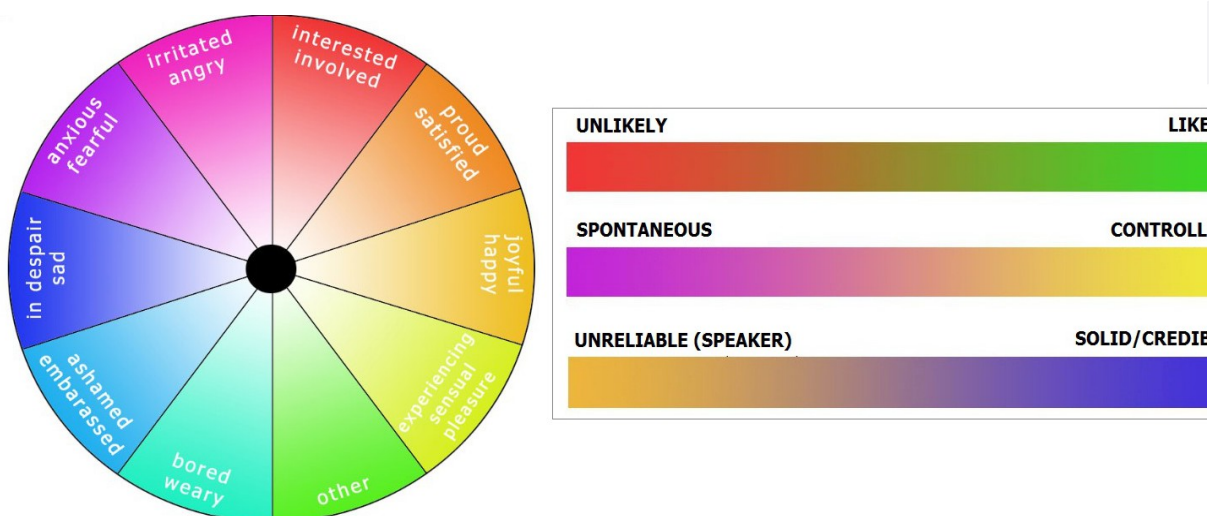
Przeprowadzono również wstępną anotację jakości głosu z wykorzystaniem graficznej reprezentacji przestrzeni cech widocznej na ilustracji 20. Na podstawie szczegółowej analizy fonetycznego opisu jakości głosu (Laver, 1980) wyróżniono następujące kategorie: *rozluźniony*, *napięty*, *przydechowy*, *szept*, *skrzeczący*, *falset*, *twardy*, *modalny*, *inne*.

Opis stanu mówcy (emocje)

Przeprowadzono anotację nagrań w podzbiorze „Emo” z wykorzystaniem elementu graficznego widocznego na ilustracji 21 (lewa strona). Zastosowano następujące etykiety:

poirytowany/zły, zainteresowany/zaangażowany, dumny/usatysfakcjonowany, radosny/szczęśliwy, doznający zmysłowej przyjemności, znudzony/znużony, zawstydzony/zmieszany, w rozpaczysmutny, niespokojny/przestraszony, inne.

Portrety emocji pozyskane z nagrań aktorów należące do tej samej rodziny emocji, a różniące się jedynie intensywnością, zostały połączone we wspólne kategorie. Intensywność była wyrażana poprzez odległość pomiędzy centrum okręgu i jego obwodem.



Ilustracja 21: Reprezentacje graficzne zastosowane na potrzeby opisu stanu mówcy/emocji oraz w ramach eksperymentu percepcyjnego

Anotacja w ramach testu percepcyjnego - ocena imitacji rzeczywistych emocji

Przeprowadzono eksperyment percepcyjny z wykorzystaniem reprezentacji graficznych przedstawionych na ilustracji 21. Uczestnicy eksperymentu (13 studentów pierwszego roku na Wydziale Neofilologii UAM) zostali poproszeni o anotację nagrań z podzbioru „Emo” bazy „Paralingua” dotyczących aktorskich imitacji nagrań z telefonu alarmowego. Przedtem nagrania aktorskie zostały

przetworzone w taki sposób, aby ich brzmienie nie odbiegało znacząco od jakości nagrań telefonicznych. W pierwszym kroku uczestnicy eksperymentu anotowali nagrania z użyciem elementu graficznego przedstawionego po prawej stronie ilustracji 21 (zob. objaśnienia w poprzednim paragrafie), natomiast w drugim kroku zostali po poproszeniu o subiektywną ocenę nagrań w trzech wymiarach:

prawdopodobne-nieprawdopodobne, kontrolowane-spontaniczne, wiarygodny(mówca)-niewiarygodny.

Wyniki eksperymentu opisano w publikacji Klessa, Wagner, Oleśkowicz-Popiel (2013).

4.7 Informacja o możliwościach korzystania z programu

Następująca informacja odnośnie wykorzystania programu została zamieszczona w programie:

Program powstał w ramach Projektu Rozwojowego nr O R00 0170 12 (Narodowe Centrum Badań i Rozwoju) w Fundacji Uniwersytetu im. Adama Mickiewicza w Poznaniu. Zgodnie z umową między partnerami konsorcjum realizującego projekt (Poznańskie centrum Superkomputerowo-Sieciowe, Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Future Voice System Sp. z o.o.), po zakończeniu projektu program może być udostępniany nieodpłatnie użytkownikom zewnętrznym na potrzeby prac badawczych, do użytku niekomercyjnego.

*Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu -
Poznański Park Naukowo-Technologiczny
ul. Rubież 46, 61-612 Poznań
tel. 061 827 97 42, fax 061 827 97 41
e-mail: fundacjauam@ppnt.poznan.pl
www.ppnt.poznan.pl*

5 Cechy parajęzykowe a zastosowania praktyczne i automatyzacja procesu charakteryzacji mówcy

UWAGA: W poniższym rozdziale zawarto raport z prac realizowanych w ramach wspólnych zadań FUAM oraz PCSS. Wykorzystano zasoby należące do PCSS (modele akustyczne, sprzęt nagraniowy, korpus mowy zaszumionej).

5.1 Charakterystyka urządzeń rejestrujących i analiza ich wpływu na wyniki rozpoznawania mowy

5.1.1 Mikrofony

Baza nagrań, na podstawie której tworzone były modele akustyczne dla systemu ASR w PCSS, nagrywana była przy użyciu mikrofonów:

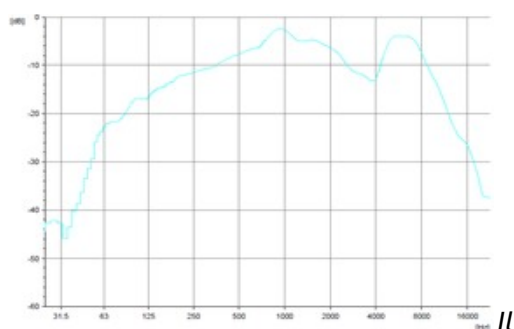
- nagłówny Sennheiser ME-3

- stołowy AKG C1000

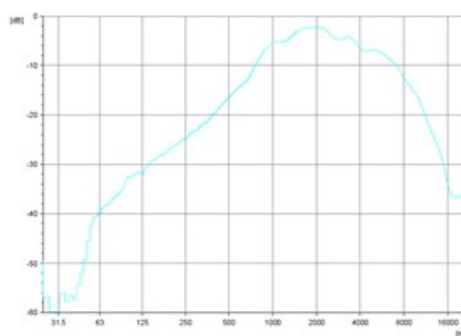
Ponieważ przyjęto założenie, że dyktowanie odbywać się będzie przy pomocy mikrofonu położonego w niewielkiej odległości od ust mówcy, do uczenia systemu wykorzystano wyłącznie część bazy nagranych mikrofonem nagłownym.

W trakcie testów, oprócz mikrofonu Sennheiser ME-3, używano również mikrofonu nagłownego Sennheiser PC26 oraz mikrofonu specjalizowanego na potrzeby dyktowania Philips SpeechMike.

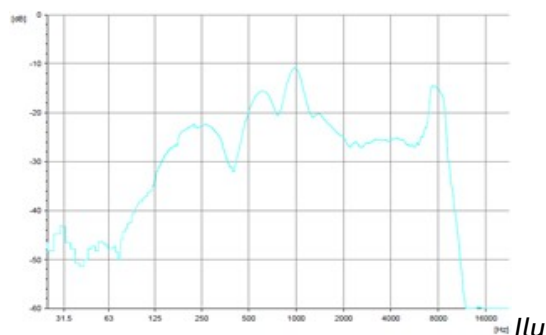
Z braku charakterystyk podanych przez producenta, zmierzono je w komorze bezchłowej przy użyciu metody MLS. Charakterystyki częstotliwościowe pokazano na ilustracjach 22, 23, 24.



Ilustracja 23: Charakterystyka częstotliwościowa - Sennheiser ME-3



Ilustracja 22: Charakterystyka częstotliwościowa - Sennheiser PC26



Ilustracja 24: Charakterystyka częstotliwościowa - Philips SpeechMike

5.1.2 Karta dźwiękowa

Sygnal z mikrofonu zamieniany jest jednokanałowo na postać cyfrową w przetworniku A/D dowolnej karty dźwiękowej komputera. Przyjęto częstotliwość próbkowania równą 16 kHz i długość próbki 16-bit. Do współpracy z mikrofonem Sennheiser ME-3 wybrano kartę zewnętrzną na złączu USB, Creative SoundBlaster Play!. Pozostałe dwa mikrofony, opisane w p.1.1, posiadają własne przetworniki A/D i są podłączane do komputera za pośrednictwem złącza USB.

Porównanie wyników adaptacji różnych modeli akustycznych oraz wpływu doboru mikrofonu na jakość rozpoznawania (20110603)

Model akustyczny	bez adaptacji do głosu mówcy					
	PC26		SB		SM	
	Acc[%]	Czas[%]	Acc[%]	Czas[%]	Acc[%]	Czas[%]
am_mfcc_nonorm_wi_8_ro100_tb350	58,05	335,79	84,57	290,4	56,84	484,02
am_1001_FM2_100_350_8_0.0_3_szczep	39,76	966,2	76,75	598,92	29,78	1160,96
am_lda_1103_szczepienie_100_350_8_0.0_3_wi_states_52	45,36	307,08	81,61	276,77	58,42	525,73
am_lda_1103_szczepienie_100_3060_8_wi_52	44,55	683,83	78,58	426,97	36,38	839,78

Model akustyczny	z adaptacją do głosu mówcy					
	PC26		SB		SM	
	Acc[%]	Czas[%]	Acc[%]	Czas[%]	Acc[%]	Czas[%]
am_mfcc_nonorm_wi_8_ro100_tb350	85,38	192,41	88,58	138,49	87,45	253,42
am_1001_FM2_100_350_8_0.0_3_szczep	66	785,1	83,46	338,54	79,65	495,96
am_lda_1103_szczepienie_100_350_8_0.0_3_wi_states_52	82,92	213,25	86,24	143,18	83,65	202,58
am_lda_1103_szczepienie_100_3060_8_wi_52	72,25	395,91	83,78	177,02	78,48	261,55

PC26 - nagrania wykonane z użyciem mikrofonu Sennheiser PC26

SM - nagrania wykonane z użyciem mikrofonu SpeechMike

SB - nagrania wykonane z użyciem mikrofonu Sennheiser ME3 + karta dźwiękowa Sound Blaster

Parametry dekodera: Tokens=5000 Delta=160 Insert Penalty=-25 Linguistic Coef=6 Observation Skip Force=0 Linguistic model weight=15 Linguistic model=test_tg_discUKN_min1_2 Version:1.0.1.15

Zbiór testowy: 13 mówców po 147 wypowiedzi (nagrania na trzech mikrofonach)

Opis użytych modeli:

- am_mfcc_norm_wi_8_ro100_tb350 - model akustyczny, parametryzacja mfcc, wektor 39 mfcc_O_D_A, 8 mikstur, trifony wewnątrzwyrazowe;
- am_1001_FM2_100_350_8_0.0_3_szczep - model oparty o parametryzację LDA ze stycznia 2010 roku; trifony wewnątrzwyrazowe, 8 mikstur; korpus uczący to suma zbiorów kobiet i mężczyzn / 2 - by wielkość modelu była taka sama jak modeli dopasowanych do płci; korpus zaszczipiony na ściszenie;
- am_lda_1103_szczepienie_100_350_8_0.0_3_wi_states_52 - model oparty o parametryzację stanowego LDA z marca 2010; trifony wewnątrzwyrazowe, 8 mikstur, długość wektora 52 cechy
- am_lda_1103_szczepienie_100_3060_8_wi_52 - model oparty o parametryzację LDA z marca 2010; trifony wewnątrzwyrazowe, 8 mikstur, długość wektora 52 cechy;
- test_tg_discUKN_min1_2 - model językowy zbudowany na znormalizowanym korpusie LEX, bigramy odciete na wartości 1, trigramy odciete na wartości 2.

5.1.3 Wpływ adaptacji na poprawność oraz czas rozpoznawania z użyciem różnych mikrofonów oraz różnych modeli akustycznych.

AM	PC26		SB		SM	
	Acc[%]	T[%]	Acc[%]	T[%]	Acc[%]	T[%]
38_mfcc_lo01_wi_8_ro100_tb350*	66,2	93,25	85,8	87,23	66,2	121,19
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m350_ada pt_v2	85	91,5	88,7	76,08	87,8	114,67
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m500_ada pt_v2	86,8	101,11	89,5	81,75	88,3	126,05
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m700_ada pt_v2*	88,3	74,36	89,6	58,62	88,6	96,42
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m1000_ad apt_v2	88,4	115,87	89,6	91,2	88,6	140,57
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m1400_ad apt_v2	88,7	118,67	89,6	96,02	88,6	146,8
38_mfcc_lo01_wi_8_ro100_tb350_rt2096_m2000_ad apt_v2	88,2	120,13	89,8	97,35	88,5	147,07

* = czas rozpoznawania zanizony w odniesieniu do pozostałych wyników ze względu na użycie komputera o lepszych parametrach

PC26 - nagrania wykonane z użyciem mikrofonu Sennheiser PC26

SB - nagrania wykonane z użyciem mikrofonu Sennheiser ME3 + karta dźwiękowa Sound Blaster

SM - nagrania wykonane z użyciem mikrofonu SpeechMike

Parametry dekodera: Tokens=5000 Delta=160 Insert Penalty=-20 Linguistic Coef=9 Observation Skip Force=0 Version:1.1.0.9 Language Model Weight=15 LM=proba3_lm_all_ukn1_2

Zbiór testowy: 13 mówców po 147 wypowiedzi na trzech mikrofonach.

Opis użytych modeli:

- 38_mfcc_lo01_wi_8_ro100_tb350 - model akustyczny, parametryzacja mfcc, wektor 39 mfcc_O_D_A, 8 mikstur, trifony wewnątrzwyrazowe, filtr dolnozaporowy 1Hz; parametr m - minimalna ilość przykładów potrzebnych do oszacowania odrębnej macierzy transformacji dla danego liścia w drzewie klas; parametr rt liczba liści w drzewie klas.

Poprawność rozpoznawania %Acc i względny czas rozpoznawania dla różnych mikrofonów.

mówca/tekst	opis	Poprawność [%]								
		ME3 +SB			PC26			SpeechMike		
		Corr [%]	Ins [%]	%ACC [%]	Corr [%]	Ins [%]	%ACC [%]	Corr [%]	Ins [%]	%ACC [%]
13 osób/testset	1103 bez odcięcia	86,9	1,7	85,2	75,7	10,7	65	62,1	12,7	49,9
	1103_lp100	85,8	2,5	83,3	68,3	18	50,3	70,2	9,3	60,3
	1103_lp150	87,1	1,5	85,6	73,7	18,4	55,3	72,6	9,2	63,3

Parametry dekodera:

ASR Test Utility (C) 2010 PSNC ver. 1.0.1.17

Acoustic Model: am_mfcc_1103_szczepienie_100_350_8_0.0_3_lp100_wi / lp150_wi / wi

Recognition network: am_mfcc_1103_szczepienie_100_350_8_0.0_3_lp100_wi

Linguistic model: test_tg_discUKN_min1_2

DECODER: Tokens=2000 Delta=150 Insert Penalty=-20 Linguistic Coef=7 Observation Skip Force=0 **LM weight=15** Version:1.0

Wnioski.

Do uczenia system ASR (budowania modelu akustycznego) użyto mikrofonu Sennheiser ME-3 i ten mikrofon gwarantuje najlepszą poprawność rozpoznawania. Jednak proces adaptacji pozwala także w przypadku użycia innych mikrofonów (Sennheiser PC26 lub SpeechMike) uzyskać porównywalne wyniki.

5.2 Cechy prozodyczne i parajęzykowe mowy a charakterystyka akustyczna środowiska zewnętrznego

Sporządzono zestawienie najważniejszych cech akustycznych środowiska zewnętrznego, mogących wpływać na identyfikację cech prozodycznych mowy. Podsumowano zakresy wartości odnoszących się do natężenia, wysokości i barwy głosu, mających wpływ na stopień zrozumiałości mowy oraz wyrazistości elementów składowych wypowiedzi w języku polskim. Opisano najważniejsze czynniki powodujące zniekształcenia sygnału mowy i zmniejszające zrozumiałość mowy, ze szczególnym uwzględnieniem stosunku rodzaju i poziomu zaszumienia i hałasu, efektu Lombarda, pogłosu oraz echa, jak również zakłóceń powstających w torze elektroakustycznym. Zwrócono uwagę na sposób oddziaływania analizowanych czynników zewnętrznych na zrozumiałość i wyrazistość poszczególnych cech artykulacyjnych mowy (m.in. sposobu i miejsca artykulacji głosek), tempo mowy, natężenie dźwięku mowy i częstotliwość podstawową, iloczyn głoskowy oraz na swoistą redundancję danych w sygnale mowy, umożliwiającą percepcję dźwięku nawet w trudnych warunkach akustycznych.

Słowa wytwarzane przez mówcę muszą mieć odpowiednie natężenie, wysokość i barwę. Poziom dźwięku mowy: 30 dB - 80 dB. Średni poziom w odległości 1m: 65 dB (SPL)

5.2.1 Informacje podstawowe

Procentowy rozkład energii w poszczególnych pasmach widma mowy:

	0 Hz - 500 Hz	500 Hz - 1 kHz	1 kHz - 4 kHz
Energia mowy	65%	30%	5%
Zrozumiałość mowy	5%	30%	65%

Istotne znaczenie dla zrozumiałości mowy mają częstotliwości do 4 kHz, powyżej wpływ jest niewielki [1].

Procentowy rozkład zrozumiałości mowy w zależności od zakresu widma mowy:

Zrozumiałość mowy w %	Częstotliwość w Hz
1	62 - 125
1	125 - 250
3	250 - 500
35	500 - 1000
35	1000 - 2000
13	2000 - 4000
12	4000 - 8000

Dla rozumienia mowy polskiej najistotniejsze jest pasmo **500 - 1600 Hz**.

Percepcja dźwięków mowy nie zależy wyłącznie od ich struktury czasowo-widmowej, ale również od treści lingwistycznej.

Wyrazistość a zrozumiałość

Pojęcie wyrazistości dotyczy tych elementów fonetycznych mowy, które nie mają określonego znaczenia semantycznego (głoski, zgłoski, logatomy), natomiast zrozumiałość dotyczy elementów mowy, które mają określone znaczenie semantyczne. Zrozumiałość jest złożoną funkcją wyrazistości. Jako miarę wyrazistości lub zrozumiałości mowy przyjmuje się stosunek liczby poprawnie odebranych elementów fonetycznych do całkowitej liczby wszystkich zaprezentowanych elementów fonetycznych.

Wpływ na zrozumiałość mowy

Czynniki powodujące zniekształcenia sygnału mowy i zmniejszające zrozumiałość mowy:

5.2.2 Stosunek S/N mowy. Hałas.

Szum powoduje maskowanie lub przytłumianie sygnału głosowego. Jesteśmy w stanie akceptować dużą porcję szumów w sygnale, nim zrozumiałość spadnie znacząco, ale gdy spadek się rozpocznie, jest bardzo gwałtowny. Dla $SNR=0$ dB zrozumiałość jest na poziomie 50%, dopiero gdy $SNR > 6$ dB zrozumiałość jest zadawalająca.

Eksperymenty dotyczące zrozumiałości mowy filtrowanej pasmowo w środowisku zaszumionym, oceniane przy użyciu współczynnika artykulacji AI pokazują, że zakres częstotliwości 200 - 5000 Hz bierze największy udział w percepcji mowy. Ten zakres odpowiada częstotliwościom największej wrażliwości audytoryjnej i największej energii mowy i sugeruje, że w ewolucji języka fonemy, które są łatwo produkowane i postrzegane przez ludzkie organy przetrwały [2]. Kluczowe percepcyjne aspekty sygnału mowy są bardziej widoczne w dziedzinie częstotliwości (reprezentacja spektralna) niż czasu. Współczesne własności mowy, które są percepcyjnie istotne np. obwiednia amplitudowa (która wskazuje na strukturę sylab, rytm i prozodię) może być łatwo uzyskana ze spektrogramu. Znaczenia różnych zakresów częstotliwości mogą być mierzone poprzez to, jak negatywny wpływ na percepcję mowy ma szum. Jeśli odfiltrujemy częstotliwości poniżej 1 kHz z sygnału mowy pomyłki wystąpią w udźwięcznieniu i wpłyną negatywnie na sposób artykulacji (np. /p/ zamiast /b/ zamiast /v/). Jeśli częstotliwości powyżej 1,2 kHz wyeliminujemy z sygnału błąd dotyczyć będzie głównie miejsca artykulacji (np. /p/ zamiast /t/). Dodatkowy szerokopasmowy szum wywoła błędy udźwięcznienia, pomyłki w odbiorze spółgłosek zwarto trących a szczególnie błędy dotyczące miejsca artykulacji.

Rezultaty pokazują, że:

- dźwięczność jest postrzegana poprzez strukturę harmoniczną (silną w niskich częstotliwościach, ale obecną do ok. 3 kHz)
- rodzaj (sposób) artykulacji jest związany z niskimi częstotliwościami
- za miejsce artykulacji odpowiada rejon powyżej 1 kHz (szczególnie F2)

Sygnał mowy charakteryzuje duża grupa parametrów akustyczno-fonetycznych, przy czym nie wszystkie parametry biorą bezpośredni udział w procesie percepcji. Ta redundancja danych pozwala percypować dźwięk nawet w trudnych warunkach akustycznych. I tak obcięte szczytowe wartości sygnału mowy eliminują większość informacji amplitudowych a mimo to słuchacze z łatwością rozpoznają tekst (zachowanie tylko 1-2 % właściwych amplitud w sygnale mowy pozwala nadal rozpoznawać słowa na poziomie 80-90%).

Percepcja rytmu, intonacji, wzorów akcentowych pomaga słuchaczom zrozumieć tekst poprzez wskazywanie ważnych słów i logicznych przerw w wypowiedzi.

5.2.3 Efekt Lombarda

Efekt Lombarda - jest niezamierzoną tendencją mówiącego do zwiększenia natężenia głosu w celu poprawienia słyszalności przy mówieniu w głośnym otoczeniu. Zmiana ta obejmuje nie tylko głośność, ale także inne cechy akustyczne takie jak tonacja oraz tempo i czas trwania sylab. Jest to efekt kompensujący, który podwyższa stosunek sygnału do szumu (SNR).

Przy porównaniu mowy nagranej w towarzystwie hałasu oraz mowy nagranej w ciszy i odsłuchanej przy hałasach maskujących, ta pierwsza jest rozumiana z reguły lepiej. Różnice pomiędzy mową normalną a lombardzką zawierają:

- wzrost fonetycznej częstotliwości podstawowe F_0 ,
- przesunięcie energii z pasma niskich częstotliwości w kierunku pasm częstotliwości średnich i wysokich,
- wzrost natężenia dźwięku,
- wzrost czasu trwania samogłosek,
- nachylenie spektrum (ang. spectral tilting),
- przesunięcie centralnych częstotliwości formantów F_1 (przede wszystkim) i F_2 ,
- czas trwania słów przenoszących treść w porównaniu ze słowami funkcyjnymi jest przedłużony w większym stopniu,
- wydawniejsze ruchy mięśni twarzy, lecz nie pomagają one w takim stopniu jak ww. zmiany.

Powyższe zmiany nie mogą być kontrolowane przez nakazanie osobie mówiącej mówić tak jak w cichym otoczeniu, jednak może zostać to wyuczone.

5.2.4 Pogłos

Im dłuższy czas pogłosu, tym słabsza zrozumiałość mowy. Dlatego im dłuższy jest czas pogłosu tym wolniej powinny następować po sobie kolejne sylaby i słowa. Zwiększenie tempa mowy w warunkach pogłosowych zakłóca jej zrozumiałość [4]. Pogłos powstaje z odbić dźwięku, które powodują rozmycie dźwięku, przez co staje się on mniej czysty i wyraźny, a przez to mniej zrozumiały. Gdy opóźnienie pomiędzy falą bezpośrednią a falą odbitą jest większe niż 100 ms pojawia się wrażenie echa (występuje raczej w otwartej przestrzeni).

Czas pogłosu definiuje się najczęściej jako czas spadku poziomu natężenia dźwięku po wyłączeniu źródła o 60 dB. Tak zdefiniowany czas pogłosu oznacza się symbolem T_{60} . Czas pogłosu jest parametrem obiektywnym oceny akustycznej wnętrza, określającym ich chłonność akustyczną. Czas pogłosu zależy od wielu czynników, m.in. od: wielkości i kształtu pomieszczenia, rodzaju i rozłożenia materiałów konstrukcyjnych, częstotliwości sygnału, położenia źródła dźwięku, wilgotności i temperatury. Długość pogłosu decyduje o możliwości zastosowania pomieszczenia do różnych celów, nie daje jednak wiedzy o jakości dźwięku. Preferowane długości czasu pogłosu dla wybranych miejsc, to:

- Sala koncertowa - 1,7-2,1s (większe dla większych, mniejsze dla mniejszych zespołów)
- Opera - 1,3-1,6s (ze względu na zrozumiałość mowy)
- Studio - 0,5-1s

- Studio do nagrań lektorskich - 0,3-0,5s
- Reżysernia studia nagrań - 0,2-0,4s

Do teoretycznego wyliczania czasu pogłosu wykorzystuje się empiryczne wzory: np. wzór Sabine'a (daje wiarygodne wyniki w przypadku dużych pomieszczeń o małej i równomiernie rozłożonej chłonności akustycznej), wzór Eyringa (stosowany dla silnie wytłumionych pomieszczeń o małym czasie pogłosu), wzór Millington-Sette'a (stosowany dla wnętrz o nierównomiernym rozkładzie współczynnika chłonności akustycznej).

Schuller (2011) pokazuje, jak zmienia się poprawność rozpoznawania mowy w różnych środowiskach akustycznych - bez pogłosu i w vanie Mercedesa ($T_{30}=0,26$), pokoju gościnnym ($T_{30}=0,32$), holu ($T_{30}=1,03$), łazience ($T_{30}=0,85$), hali fabrycznej ($T_{30}=3,27$) i kościele ($T_{30}=3,88$). Odpowiednio używając do testów bazy słownej spontanicznej (TUM AVIC zaprezentowanej na INTERSPEECH 2010, emocjonalnej bazy eNTERFACE z 2006 oraz aktorskiej, berlińskiej EMO-DB) uzyskano średnie wartości poprawności rozpoznawania niższe o 29,8% w pogłosie w stosunku do prezentacji dźwięków w komorze bezechowej (bez pogłosu). Schuller podaje także pewne informacje o zmianach w podstawowych parametrach akustycznych mowy: obwiedni widma, energii, F_0 , amplitudzie, szerokości i częstotliwości formantów 1-5, MFCC 1-16, shimmer+Ritter+HNR dla dźwięku „czystego” i wymienionych wcześniej 6 pomieszczeń.

5.2.5 Zniekształcenia toru akustycznego.

Możemy do nich zaliczyć następujące zniekształcenia oryginalnego sygnału powodowane przez tor elektroakustyczny:

- Zniekształcenia liniowe - nierównomierność charakterystyki częstotliwościowej, wzmocnienie lub tłumienie sygnału, czyli przesunięcie jego dynamiki w górę lub w dół
- Zniekształcenia nieliniowe - modulacja, demodulacja, kompresja, ekspansja (te 2 ostatnie przypadki wiążą się z przekształceniem objętości informacyjnej), kwantyzacja, próbkowanie, dekodowanie.
- Zniekształcenia fazowe, amplitudowe (np. obcięcie sygnału, ograniczenie wartości chwilowej sygnału, ograniczenia czasowe, korekcja zależności fazowych między poszczególnymi parametrami sygnału).

5.3 Rozpoznawanie mowy/mówcy w warunkach zaszumionych

Rozpoznawanie mowy w warunkach zaszumionych wiąże się ze zwiększeniem stopnia trudności tego zadania, w związku koniecznością uwzględnienia

dodatkowych złożonych czynników wpływających na jakość rozpoznawania. W sytuacji, gdy rozpoznawanie ma działać dla różnego rodzaju głosów oraz gdy nie jest możliwa pełna przewidywalność możliwych rodzajów i poziomów zakłóceń zewnętrznych, należałoby zwrócić szczególną uwagę nie tylko na charakterystykę szumów, ale również na jej relację względem charakterystyki głosu ludzkiego, zarówno cech związanych z systemem językowych, specyfiką głosek, jak również indywidualnych cech mówcy.

5.3.1 Testy rozpoznawania mowy z wykorzystaniem korpusu mowy zaszumionej

Przeprowadzono testy rozpoznawania mowy z wykorzystaniem korpusu mowy zaszumionej zarejestrowanego w ramach obecnego projektu w PCSS (zob. raport PCSS odnośnie szczegółów dotyczących bazy). Do testu użyto nagrania dla 18 mówców, z których każdy dostarczył realizację 84 komend głosowych. Testy wykonano z wykorzystaniem systemu ARM w wersji 89. Do testów wykorzystano uproszczoną sieć rozpoznawania (zawierającą jedynie komendy).

W tabeli poniżej przedstawiono wyniki testów rozpoznawania z wykorzystaniem korpusu mowy zaszumionej w pojeździe z różną konfiguracją odnośnie charakterystyki zakłóceń (włączona/wyłączona klimatyzacja, otwarte/zamknięte okno). Każda konfiguracja została przetestowana z użyciem dwóch różnych mikrofonów (ME3, PC26).

Wyniki uporządkowano w kolejności od najlepszych do najgorszych wyników:

Level	Acc [%]					
	sil		airc		wind	
	ME3	PC26	ME3	PC26	ME3	PC26
The highest	93.00	93.60	92.80	92.70	92.90	93.30
Higher	92.80	93.30	92.40	92.60	92.80	93.20
High	92.50	93.20	92.20	92.30	92.30	93.00
Mean	92.30	92.90	92.30	92.10	92.40	92.80
Low	92.20	92.80	92.30	92.00	92.30	92.50
Lower	92.30	92.70	92.10	91.70	92.20	92.50
The lowest	92.20	92.50	91.90	91.30	92.10	92.20

Objaśnienie skrótów w tabeli:

sil - wyłączona klimatyzacja, zamknięte okno (nawiew)

airc - włączona klimatyzacja, zamknięte okno

wind - wyłączona klimatyzacja, otwarte okno obok pasażera (mówcy)

Level - poziom jakości rozpoznawania (od najniższej do najwyższej) - predefiniowane poziomy jakości rozpoznawania w systemie ARM

5.4 Analiza anotacji korpusu mowy zaszumionej i mowy pod wpływem alkoholu

Korpus mowy zaszumionej oraz korpus mowy pod wpływem alkoholu PCSS anotowano manualnie na poziomie ortograficznym i automatycznie na poziomie fonetycznym. Jako narzędzie zastosowano program *Transcriber*. Użyto specyfikacji anotacji opartej na specyfikacji SPEECON, użytej również wcześniej do anotacji bazy JURISDIC, co jest w pełni uzasadnione ze względu na utrzymanie ciągłości badań prowadzonych nad rozwojem narzędzi, modeli oraz systemów rozpoznawania mowy/mówców. Specyfikacja w przypadku bazy mowy pod wpływem alkoholu została poszerzona o oznaczenia związane z moą słyszaną w tle [bspeech] - „background speech” (w bazie mowy alkoholowej zarejestrowano nagrania dialogów, w związku z czym, poza głosem głównego mówcy słyszalny jest w nich również głos osoby przeprowadzającej rozmowę).

Być może w przyszłości, poza tradycyjną anotacją nagrań z zastosowaniem etykiet dla stosunkowo ograniczonej liczby zjawisk pozajęzykowych oraz zakłóceń (wynikającej ze specyfikacji SPEECON), przydatne byłoby również przeprowadzenie szerszej zakrojonych analiz z wykorzystaniem anotacji cech na skali ciągłej lub na płaszczyźnie, np. za pomocą specjalnie zaprojektowanej reprezentacji graficznej przestrzeni cech w programie Annotation System (zob. rozdz. 4 wyżej). Celem byłaby m.in. weryfikacja kategoryzacji zjawisk pozajęzykowych przyjmowanych w istniejących specyfikacjach anotacji. Analizy tego rodzaju mogłyby zatem dotyczyć np. percepcyjnej weryfikacji:

- oceny zrozumiałości mowy
- oceny poziomu i rodzaju zakłóceń
- oceny wpływu zakłóceń na zrozumiałość mowy oraz identyfikowalność mówcy
- oceny stanu mówcy (nacechowania afektywnego czy też bezpośrednio subiektywnej oceny trzeźwości na podstawie brzmienia głosu)
- identyfikacji mówcy na podstawie odsłuchu nagrań

Reprezentacja graficzna mogłaby być zaimplementowana jako zestaw niesymetrycznych skal (min-max), na których anotujący (lub później uczestnik eksperymentu percepcyjnego) oznaczałby odpowiedzi. Jako podstawa do oceny stanów afektywnych może służyć płaszczyzna valence-potency (z ewe)

Przeprowadzenie szeregu takich analiz, a następnie eksperymentów percepcyjnych mogłoby dostarczyć dodatkowych informacji odnośnie możliwych uogólnień dotyczących charakterystyki indywidualnych cech głosu ludzkiego i jego zależności od czynników zewnętrznych.

6 Prezentacja i popularyzacja wyników i tematyki badań (publikacje, prezentacje prac, udział w konferencjach)

6.1 Publikacje

- Karpinski, M. (2012) The boundaries of language: Exploring paralinguistic features. *Lingua Posnaniensis*, LIV (2), pp. 37-54. DOI: 10.2478/v10122-012-0013-1
- Karpiński, M. Acoustic features of filled pauses in Polish task-oriented dialogues. (przyjęte do druku w: *Archives of Acoustics*)
- Klessa, K. (2011/2012) Polish segmental duration: selected observations based on corpus data [w:] *Speech and Language Technology*, Red. Gibbon, Dafydd, Hirst, Daniel, Campbell, Nick, Poznan, PTFon, Vol. 14/15 (2011/2012), pp. 94-104.
- Klessa, K. (2013) Annotation of paralinguistic features for the needs of speaker characterisation. (przyjęte do druku w: *Proceedings of XI Nordic Prosody Conferences*, Wyd. Peter Lang).
- Klessa, K. (2013) Computer-assisted analysis of scalar features in spoken language corpora (przyjęte do druku w: *Proceedings of 19th International Congress of Linguists*, Geneva, Switzerland).
- Klessa, K., Karpiński, M. (2012) Annotating paralinguistic features in quasi-spontaneous speech. Adding the "vision" component? [w:] *Proceedings of the Workshop on Vision and Language, December 13th and 14th, 2012*, Sheffield, UK.
- Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M. (2013) Using "Paralingua" database for investigation of affective states and paralinguistic features (w przygotowaniu, przedłożone do druku w: *Speech and Language Technology* Vol. 16), 2013.
- Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M., Karpiński, M. (2013) "Paralingua" - a new speech corpus for the studies of paralinguistic features (przyjęte do druku w: *Proceedings of 5th International Conference on Corpus Linguistics*, Alicante, Spain), 2013.
- Wagner, A. (2011/2012) Automatic labeling of prosody [w:] *Speech and Language Technology*, Red. Gibbon, Dafydd, Hirst, Daniel, Campbell, Nick, Poznan, PTFon, Vol. 14/15, pp. 157-167.
- Wagner, A. (2012) Speech rhythm in native and non-native Polish. [w:] *Proceedings of International Conference of Experimental Linguistics: ExLing 2012. 27-29 August 2012*, Athens, Greece.
- Wagner, A. *Analysis of speech rhythm in native and non-native Polish using rhythm metrics*, (w przygotowaniu, przedłożone do recenzji).

6.2 Prezentacje konferencyjne

- Karpiński, M., Klessa, K. (2012) *Paralinguistic features of utterances: Exploring the boundaries of language*. Przyjęto do prezentacji podczas międzynarodowej konferencji pn. *Talk, Action, Interaction*, w dn. 3-5.10.2012 w Łodzi.
- Karpiński, M., Klessa, K. (2012) *Paralinguistic features of utterances: Implementing categorisations and levels*. Referat wygłoszono podczas warsztatów naukowych w ramach walnego zjazdu Polskiego Towarzystwa Fonetycznego pt. *Rozpoznawanie mowy i mówcy*, Poznań.
- Klessa, K. (2012) *Annotation of paralinguistic features for the needs of speaker characterisation*. Referat przedstawiono w ramach międzynarodowej konferencji dotyczącej prozodii języków wokół M. Bałtyckiego *XI Nordic Prosody Conference*, 15-17.08.2102 University of Tartu, Estonia.
- Klessa, K. (2013) Computer-assisted analysis of scalar features in spoken language corpora. Przyjęto do wygłoszenia podczas: *19th International Congress of Linguists*, 22-27 lipca 2013 r., Geneva, Switzerland.
- Klessa, K., Karpiński, M. (2012) *Annotating paralinguistic features in quasi-spontaneous speech. Adding the "vision" component?* referat przedstawiono w ramach międzynarodowych warsztatów *Vision & Language* organizowanych wspólnie w Sheffield w Wlk. Brytanii przez University of Brighton oraz University of Sheffield.
- Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M., Karpiński, M. (2013) "Paralingua" - a new speech corpus for the studies of paralinguistic features. Przyjęto do wygłoszenia podczas *5th International Conference on Corpus Linguistics*, 14-16 marca 2013, Alicante, Spain.
- Wagner, A. (2012) Kolokwium naukowe *Phonetisches Kolloquium*, Institut für Phonetik, Universität des Saarlandes, Niemcy, 28 listopada 2012. (referat wygłoszony na zaproszenie organizatorów).
- Wagner, A. (2012) *Speech rhythm in native and non-native Polish*. Referat przedstawiono w ramach międzynarodowej konferencji ISCA: *International Conference of Experimental Linguistics: ExLing 2012*. 27-29 August 2012, Athens, Greece.
- Wagner, A. (2012) *Analysis of rhythm in native and non-native Polish using rhythm metrics*. Przyjęto do prezentacji podczas warsztatów naukowych w ramach walnego zjazdu Polskiego Towarzystwa Fonetycznego pt. *Rozpoznawanie mowy i mówcy*.

6.3 Prace magisterskie

Prace magisterskie powiązane z tematyką prac w ramach projektu (opiekun prac: dr Katarzyna Klessa). Tematy:

- Analiza doboru środków leksykalnych w nagraniach mowy quasi-spontanicznej.

- Analiza struktur składniowych w nagraniach mowy quasi-spontanicznej.
- Rodzaje i liczba pauz wypełnionych w nagraniach mowy quasi-spontanicznej.

7 Literatura przedmiotu

- Adami, A. 2005. Prosodic Modeling for Speaker Recognition Based on Sub-Band Energy Temporal Trajectories. *Proceedings of IEEE Conference 2005*.
- Agrawal, S., Shruti, A.K., Rama Krishna, C. 2010. Prosodic feature based text-dependent speaker recognition using machine learning algorithms. *International Journal of Engineering Science and Technology*, Vol. 2(10), pp. 5150-5157
- Ainsworth, W. A. 1997. Some approaches to automatic speech recognition. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*,. Oxford: Blackwell Publ, pp. 721-743.
- Amerman, J. D. and Parnell, M. M. 1992. Speech timing strategies in elderly adults. *Journal of Voice*, 20, pp. 65-67.
- Amino K., Sugawara, T., Arai T., 2006. Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties, *Acoust. Sci. & Tech.* 27(4).
- Atal, B.S. 1972. Automatic speaker recognition based on pitch contours, *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 1687-1697.
- Atal, B.S. 1974. Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America*, Vol. 55, No. 6, pp. 1304-1312, 1974.
- Atal, B.S. 1976. “Automatic Recognition of Speakers from Their Voices,” *Proceedings of the IEEE* Vol. 64, pp. 460-475, 1976.
- Attili, J., M. Savic, and J. Campbell. “A TMS32020-Based Real Time, Text-Independent, Automatic Speaker Verification System,” In *International Conference on Acoustics, Speech, and Signal Processing in New York*, IEEE, pp. 599-602, 1988.
- Banse, R.; Scherer, K. R. (1996) "Acoustic profiles in vocal emotion expression". *Journal of Personality and Social Psychology*, 70(3), pp. 614-636.
- Basseville, M. “Distance measures for signal processing and pattern recognition,” *Signal Process.*, vol. 18, pp. 349-369, 1989.
- Blatchford, H., Foulkes, P., 2006, “Identification of voices in shouting”, The International Journal of Speech, Language and the Law, Vol 13, p 241-254
- Bockle, "The SRI NIST 2008 Speaker Recognition Evaluation System", ICASSP, Taipei, Taiwan, 2009.
- Bonastre, J.F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin G., Evans N., 2007 “ALIZE/ SpkDet: a state-of-the-art open source software for speaker recognition” proceeding Interspeech 2007.

- Brunet, P. M., Charfuelan, M., Cowie, R., Schröder, M., Donnan, H., & Douglas-Cowie, E. 2010. Detecting politeness and efficiency in a cooperative social interaction. In *Proc. Interspeech*. Makuhari, Japan.
- Campbell, J.P. 1995. Testing with The YOHO CD-ROM Voice Verification Corpus, In *International Conference on Acoustics, Speech, and Signal Processing in Detroit*, IEEE, pp. 341-344, 1995. Available: <http://www.biometrics.org/>.
- Campbell, J.P. 1997. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462.
- Campbell, J.P., T. E. Tremain, and V. C. Welch. The Federal Standard 1016 4800 bps CELP Voice Coder, *Digital Signal Processing*, Vol. 1, No. 3, pp. 145 -155, 1991.
- Campbell, W.M., Z. N. Karam, D. E. Sturim 2009. Speaker Comparison with Inner Product Discriminant Functions in NIPS 2009, pp. 364-37.
- Charfuelan, M., & Schröder, M. 2011. Investigating the prosody and voice quality of social signals in scenario meetings. *Proc. Affective Computing and Intelligent Interaction*. Memphis, TN, USA.
- Charfuelan, M., & Schröder, M. 2011. The vocal effort of dominance in scenario meetings. *Proc. Interspeech*. Florence, Italy.
- Charfuelan, M., Schröder, M., & Steiner, I. 2010. Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In *Proc. Interspeech*. Makuhari, Japan.
- Che, C. & Lin, Q. 1995. "Speaker recognition using HMM with experiments on the YOHO database," In *EUROSPEECH in Madrid*, ESCA, pp. 625- 628.
- Colombi, J., D. Ruck, S. Rogers, M. Oxley, and T. Anderson, 1996. Cohort selection and word grammar effects for speaker recognition, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996, pp. 85-88.
- Dau, T., D. Puschel, and A. Kohlrausch, "A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, 3615-3622, 1996.
- Davis, S.B. and Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 28, no. 28, pp. 357-366, Aug. 1980.
- Deller jr., J.R., John H. L. Hansen and J G Proakis 2000. *Discrete-Time Processing of Speech Signal*, 2nd ed. New York: IEEE Press.
- Devijver, P.A. "On a new class of bounds on Bayes risk in multihypothesis pattern recognition," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 70-80, 1974.
- Devillers, L., Vasilescu, I. 2003. Prosodic cues for emotion characterization in real-life spoken dialogs. In *ISCA Eurospeech*, Geneva, September 2003.
- Doddington, G.R. 1985. "Speaker recognition—Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651-1664, Nov. 1985.

- Matrouf, D., Bonastre, J-F., Fredouille, C., Larcher, A., Mezaache, S., McLaren, M., Huenupan, F. 2008. LIA GMM-SVM system description: NIST SRE in NIST SRE, Montreal (Canada).
- Duda, R. & Hart, P. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.
- Farrus, M. and Hernando, J. 2009. Using jitter and shimmer in speaker verification, IET signal proc., vol. 3, no. 4, pp. 247-257, Nov 2009.
- Fauve, B, Evans, N, Mason, J, Improving the performance of textindependent short duration SVM- and GMM-based speaker verification, in Proc. Odyssey: the Speaker and Language Recognition Workshop, 2008
- Flanagan, J. 1972. *Speech Analysis Synthesis and Perception*. 2nd ed., Berlin: Springer-Verlag.
- Furui, S. 1981. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 2 , pp. 254-272.
- Furui, S. 1991. Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques." *Speech Communication*, Vol. 10, pp. 505-520.
- Furui, S. 1997. Recent Advances in Speaker Recognition," *Pattern Recognition Letters*, Vol. 18, pp. 859-872.
- Gish, H. & Schmidt, M. 1994. Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, Vol. 11, No. 4, pp. 18-32, 1994.
- Gnanadesikan, R. and Kettenring, J. R., 1989. Discriminant analysis and clustering, *Statistical Sci.*, vol. 4, no. 1, pp. 34-69.
- Grimaldi, M. and Cummins, F. 2008. Speaker Identification Using Instantaneous Frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1097-1111, 2008.
- Hall, J. J. and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Patt. Anal. and Mach. intell.*, vol. 16, pp. 66-75, Jan. 1994.
- Hammerschmidt K., Jürgens U. (2007) "Acoustical correlates of affective prosody". *Journal of Voice*, 21(5), pp. 531-40.
- Hermansky, H. 1990. "Perceptual linear prediction (PLP) analysis of speech", *JASA*, Vol. 87, No. 4 April 1990.
- Higgins, A. 1990. "YOHO Speaker Verification," *Speech Research Symposium*, Baltimore, 1990.
- Higgins, A., and Wohlford, R. E. 1986. A new method of text independent speaker recognition, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, pp. 869-872.
- Higgins, A., L. Bahler, and J. Porter 1991. Speaker Verification Using Randomized Phrase Prompting, *Digital Signal Processing*, Vol. 1, No. 2 , pp. 89-106.

- Higgins, A., L. Bahler, and J. Porter 1991. Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 2, pp. 89-106.
- Higgins, A., L. Bahler, and J. Porter, 1993. Voice Identification Using Nearest Neighbor Distance Measure," In *International Conference on Acoustics, Speech, and Signal Processing in Minneapolis*, IEEE, pp. 375-378.
- Higgins, A.L. and R. E. Wohlford 1986. A New Method of Text-Independent Speaker Recognition," In *International Conference on Acoustics, Speech, and Signal Processing in Tokyo*, IEEE, pp. 869-872.
- Hojan, E. – Dopasowanie aparatów słuchowych, MEDITON, Łódź 2009.
- Hojan, E. – Zasady nagłaśniania pomieszczeń i przestrzeni otwartej, WN UAM, Poznań 1988.
- Hollien, H., Doherty, E.T. 1982. Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*. 10.p.139-148.
- Itakura, F. 1975. Line Spectrum Representation of Linear Predictive Coefficients," *Transactions of the Committee on Speech Research, Acoustical Society of Japan*, Vol. S75, No. 34, 1975.
- Itakura, F., 1975. "Line spectrum representation of linear predictive coefficients," *Trans. Committee Speech Research, Acoustical Soc. Japan*, vol. S75, p. 34, 1975.
- Jacewicz, E., Fox, R., O'Neill, C. 2009. *Language Variation and Change*, 21 (2009), 233-256. Printed in the U.S.A. CUP.
- 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*. Vol. 8 of *Research in Text Theory*. Walter de Gruyter, Berlin, pp. 203-225.
- Kailath, T., 1967. The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans. Commun. Technol.*, vol. COM-15, no. 1, pp. 52-60, 1967.
- Kajarekar, S, 2008. Phone-based Cepstral Polynomial SVM System for Speaker Recognition, proceeding of Interspeech 2008
- Kang, G. and L. Fransen, *Low Bit Rate Speech Encoder Based on Line-Spectrum-Frequency*, NRL, NRL Report 8857, 1985.
- Kreiman J. and Gerrat B. R. 2005. Perception of aperiodicity in pathological voice, *J. Acoust. Soc. Amer.*, vol. 117, pp. 2201-2211.
- Lamel, L, Gauvain, J.L. 2000. Speaker verification over the telephone. *Speech Communication* 31(2-3): pp.141-154.
- Laver, J. (1980, wyd. 2009) *The phonetic description of voice quality*. Cambridge Studies in Linguistics. Cambridge University Press.
- Lee, Y.-T. 1991. Information-theoretic distortion measures for speech recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 330-335, 1991.

- Lei, H. & E. Lopez "Mel, Linear, and Antimel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition", *Interspeech 2009*, Brighton, UK, 2009.
- Li, H. et al. 2009. "The I4U System In NIST 2008 Speaker Recognition Evaluation", *ICASSP*, Taipei, Taiwan, 2009.
- Li, K. P. and Wrench, Jr., E. H. 1983. Text-independent speaker recognition with short utterances, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Boston, MA, 1983, pp. 555-558.
- Li, Q. 2009. An auditory-based transform for audio signal processing: Applications of Signal Processing to Audio and Acoustics, 2009, WASPAA '09.
- Liscombe J., Venditti J., Hirschberg J. 2003. "Classifying subject ratings of emotional speech using acoustic features". In: *Proceedings of Interspeech'2003*.
- Lu, X., Dang, J. Physiological feature extraction for textindependent speaker identification using non-uniform subband processing, *Proc. of ICASSP*, 2007.
- Makhoul, J. 1975. "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol. 63, pp. 561-580, 1975.
- Mammone, R., X. Zhang, and R. Ramachandran, "Robust Speaker Recognition-A Featurebased Approach," *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 58-71, 1996.
- Mammone, R., X. Zhang, and R. Ramachandran, "Robust speaker recognition—A feature-based approach," *IEEE Signal Processing Mag.*, vol. 13, no. 5, pp. 58-71, 1996.
- Markel, J.D. & Davis, S. B. 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 1, pp. 74-82.
- Markel, J.D. and S. B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base," *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-27, No. 1, pp. 74-82, 1979.
- Martin, A. and M. Przybicki, 1997. 1997 Speaker Recognition Evaluation," In *Speaker Recognition Workshop*, editor A. Martin (NIST). Section 2. Maritime Institute of Technology, Linthicum Heights, Maryland, June, pp. 25-26.
- Martin, A., G. Doddington, T. Kamm, M. Ordowski, and M. Przybicki, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. on Speech Communication Technology*, Rhodes, Greece, vol. 4, 1997, pp. 1895-1898.
- Mashao, D. J. and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, pp. 147-155, Jan. 2006.
- McGhee, F., 1937. "The reliability of the identification of voice", *J.Gen.Psychol.*, n° 17, p249-271.
- Mixdorf, H., Pfizinger, H. R., Grauwinkel, K., 2005. Towards objective measures for comparing speaking styles. In: *Proc SPECOM*. Patras, Greece, pp. 131-134.

- J. M. Montero , J. Gutiérrez-arriola , J. Colás , E. Enríquez , J. M. 1999. "Analysis and modelling of emotional speech in Spanish". In: Proc. of ICPhS'99.
- Mozziconacci, S.J.L and D.J. Hermes 1999. "Role of intonation patterns in conveying emotion in speech", ICPhS99, San Francisco, pp. 2001-2004.
- Murthy, K. S. R. and B. Yegnanarayana, 2008. "Epoch extraction from speech signal," IEEE Trans. Audio Speech and Language Process., vol. 16, no. 8, pp. 1602-1613, November 2008.
- Naik, J. 1990. Speaker verification: A tutorial, *IEEE Commun. Mag.*, vol. 28, pp. 42-48.
- O'Shaughnessy, D. 1987. Speech communication, human and machine," *Digital Signal Processing*. Reading, MA: Addison-Wesley, 1987.
- O'Shaughnessy, D. 1987. *Speech Communication, Human and Machine*. Digital Signal Processing, Reading: Addison-Wesley.
- Oppenheim, A.V. and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- Pammi, S., & Schröder, M. (2011). Evaluating the meaning of synthesized listener vocalizations. Proc. Interspeech. Florence, Italy.
- Papcun, G. 1997. Commensurability Among Biometric Systems: How to Know When Three Apples Probably Equals Seven Oranges, In *Proceedings of the Biometric Consortium, 9th Speaker Recognition 25 Meeting*, editor J. Campbell (NSA). Holiday Inn, Crystal City, Virginia. (drugie wydanie?)
- Papcun, G. 1974. Commensurability among biometric systems: How to know when three apples probably equals seven oranges," in *Proc. Biometric Consortium, 9th Meeting, J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304-1312.
- Parsons, T. 1987. *Voice and Speech Processing, Communications and Signal Processing*, S. Director, Series Ed. New York: McGraw-Hill.
- Pati, D. and Prasanna, S. R. M. (w druku?) Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information, Accepted for publication, *Int. J. of Speech Technology*, Springer.
- Pati, D. and Prasanna, S. R. M. 2010. Speaker recognition from excitation source perspective, *IETE Technical Review*, vol. 27, no. 2, pp. 138-157, Mar 2010.
- Pati, D. and Prasanna, S. R. M. Processing of linear prediction residual in spectral and cepstral domains for speaker information," Communicated to *Int. J. of Speech Processing*, Springer.
- Pati, D. and Prasanna, S. R. M. Speaker information from sub-band energies of linear prediction residual, in *Proc. NCC 2010*, pp. 1-4.
- Pentz, A. 1990. "Speech science (SPATH 4313) class notes," Oklahoma State University, Stillwater.

- Plumpe, D. "Modeling of the glottal flow derivative waveform with application to speaker identification," M.S. thesis, Massachusetts Institute of Technology, Cambridge, 1997.
- Pruzansky, S. and Mathews, M. V. 1964. Talker-Recognition procedure based on Analysis of Variance," J. Acoust. Soc. Amer., vol. 36, no. 11, pp. 2041-2047.
- Przybocky, M. and Martin, A. 2000. The NIST-1999 speaker recognition evaluation- An overview, Digital signal processing, vol. 10, pp. 1-18, 2000.
- Przybocky, M. and Martin, A. 2003. NIST speaker recognition evaluation plan, in Proc. NIST Speaker Recognition Workshop, College Park, MD.
- Quatieri, T.F., N. Malyska, and D.E. Sturim. "Auditory Signal Processing As A Basis For Speaker Recognition", 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 19-22 Oct. 2003.
- Quene, H. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *J. Acoust. Soc. Am.* 123, 2.
- Rabiner, L. and B. H. Juang, *Fundamentals of Speech Recognition*, Signal Processing, editor A. Oppenheim. Englewood Cliffs: Prentice-Hall, 1993.
- Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.
- Rabiner, L.R. and B.-H. Juang, *Fundamentals of Speech Recognition*, Signal Processing, A. Oppenheim, Series Ed. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- Rabiner, L.R. and R. Schafer, *Digital Processing of Speech Signals*, Signal Processing, A. Oppenheim, Series Ed. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- Reynolds, D. 1995. Speaker identification and verification using Gaussian mixture speaker models, *Speech Commun.*, vol. 17, pp. 91-108, 1995.
- Reynolds, D. and B. Carlson 1995. Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers," in *Proc. EUROSPEECH*, Madrid, Spain, pp. 647-650.
- Reynolds, D. and R. Rose 1995. Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83.
- Reynolds, D.A. 1994. Experimental evaluation of features for robust speaker identification, *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639-643.
- Rosenberg, A. 1976. Automatic Speaker Verification: A Review," *Proceedings of the IEEE*, Vol. 64, No. 4, pp. 475-487.
- Rosenberg, E. and F. K. Soong, "Recent Research in Automatic Speaker Recognition," In: *Advances in Speech Signal Processing*, ed. S. Furui and M. M. Sondhi. pp. 701-738. New York: Marcel Dekker, 1992.

- Rosenberg, E. J. DeLong, C-H. Lee, B-H. Juang, and F. K. Soong 1992. The Use of Cohort Normalized Scores for Speaker Verification," In *International Conference on Spoken Language Processing in Banff*, University of Alberta, pp. 599-602.
- Saito, S. and K. Nakata 1985. *Fundamentals of Speech Signal Processing*. Tokyo: Academic Press.
- Sakoe, H. and S. Chiba 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-26, No. 1, pp. 43-49.
- Sapożkow, M.A. - Sygnał mowy w telekomunikacji i cybernetyce, WNT Warszawa 1966.
- Saratxaga I., Navas E., Hernáez I., Luengo I. (2006) "Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque". In: *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*.
- Saslove, H., Yarmey, A.D, 1980. Long-term Auditory Memory: Speaker Identification", *Journal of Applied Psychology*, Vol. 65, p111-116.
- Schiel, F. and Heinrich, C. 2009. Laying the foundation for in-car alcohol detection by speech. In *Proc. INTERSPEECH 2010*, Brighton, UK, pp. 983-986,
- Schwartz, R. S. Roucos, and M. Berouti, "The Application of Probability Density Estimation to Text Independent Speaker Identification," In *International Conference on Acoustics, Speech, and Signal Processing in Paris*, IEEE, pp. 1649-1652, 1982.
- Shaughnessy, D.O. 1986. Speaker recognition, vol. 3, Oct. 1986, pp. 4-17.
- Scherer K. 2003. "Vocal communication of emotion: A review of research paradigms". *Speech Communication*, Volume 40, pp.227-256.
- Schmidt, T., Elenius, K. 2010. Multimedia Encoding and Annotation, In: *Common Language Resources and Technology Infrastructure. Interoperability and Standards*. CLARIN Deliverable D5.C-3 (pp.123-145).
- Schuller, B. 2011. *Affective speaker state analysis in the presence of reverberation*, *Int J Speech Technol* (2011) 14: 77–87.
- Smith, Steven W. 1997. *The Scientist's and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, San Diego, California.
- O'Shaughnessy, D. – *Speech Communication Human and Machine*, Addison-Wesley PC, 1987.
- Soong, F., A. Rosenberg, L. Rabiner, and B-H. Juang, 1985. A Vector Quantization Approach to Speaker Recognition," In *International Conference on Acoustics, Speech, and Signal Processing in Florida*, IEEE, pp. 387-390.
- Soong, F.K., A. E. Rosenberg, L. R. Rabiner, and B. H. Juang 1987. "A Vector Quantization Approach to Speaker Recognition." *AT&T Technical Journal*, Vol. 66, No. 2 , pp. 14-26.

- Sturim, D.E., W. M. Campbell, Z. N. Karam, D. A. Reynolds, F. S. Richardson "The MIT Lincoln Laboratory 2008 Speaker Recognition System", Interspeech, Brighton, UK, 2009.
- Sutherland, A., and M. Jack, "Speaker Verification." In *Aspects of Speech Technology*, editors M. Jack and J. Laver, Edinburgh: Edinburgh University Press, pp. 185-215, 1988.
- Tishby, N.Z. "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 39, No. 3, pp. 563 - 570, 1991.
- Trouvain, J. & Schröder, M. 2004. How (Not) to Add Laughter to Synthetic Speech *Proc. Workshop on Affective Dialogue Systems* Kloster Irsee, Germany, pp. 229-232.
- Truong, Khiet P. and Trouvain, J. 2012. Laughter annotations in conversational speech corpora - possibilities and limitations for phonetic analysis. In: 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012), 26 May 2012, Istanbul, Turkey.
- Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M. & Barry, W. J. 2006. Modelling personality features by changing prosody in synthetic speech. *Proc. Speech Prosody 2006*, Dresden, Germany.
- Van Dommelen, W.A., 1987, The contribution of speech rhythm and pitch to speaker recognition", *Language and Speech*, Vol. 30, Part 4, pp. 325-338.
- Ververidis D. & Kotropoulos C. 2006. "Emotional speech recognition: Resources, features, and methods". *Speech Communication*, Vol. 48, No. 9., pp. 1162-1181.
- Ververidis D., Kotropoulos C., Pitas I. 2004. "Automatic emotional speech classification". In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '04*.
- Wada, xxx Takahiro Shinozaki, and Sadaoki Furui (Tokyo Institute of Technology) "Investigations of features and estimators for speech-based age estimation *Proc. of the Second APSIPA Annual Summit and Conference*, pp.470-473, (2010-12)
- Wolf, J. J., 1972. Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.* vol. 51, no. 2, pp. 2044-2055.
- Yapanel, U. and Hansen, J. A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition., *Speech Communication*, Vol. 50, 2, Feb. 2008.
- Zheng, N., Lee, T., and Ching, P. C. 2007. Integration of complimentary acoustic features for speaker recognition," *IEEE signal process. Lett.*, vol. 14, no. 3, pp. 181-184, March 2007.