FUNDACJA UNIWERSYTETU
IM. A. MICKIEWICZA W POZNANIU

RAPORT merytoryczny szczegółowy nr 1

opracowania z roku 2011

Autorzy raportu i wykonawcy opisanych prac: Katarzyna Klessa (główne zadania: koordynacja działań zespołu, bazy danych, narzędzia), Maciej Karpiński (główne zadania: analiza cech pozajęzykowych w mowie), Agnieszka Wagner (główne zadania: analiza emocji w mowie)

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

I. Cechy para- i pozajęzykowe wypowiedzi: definicje i inwentarze

II. Emocje w mowie - współczesne standardy i parametryzacje

III. Przegląd baz danych do zastosowań w rozpoznawaniu, weryfikacji lub identyfikacji mówców

Spis treści

I. Cechy para- i pozajęzykowe wypowiedzi: definicje i inwentarze

W poniższym raporcie:

1. Zebrano informacje na temat sposobu rozumienia i definiowania cech parajęzykowych[1] wypowiedzi.
2. Sformułowano roboczo definicję cech parajęzykowych w kontekście bieżącego projektu
3. Zgromadzono reprezentatywną próbę inwentarzy cech parajęzykowych zastosowanych w wybranych badaniach, w szczególności w opisie korpusów językowych i studiach nad identyfikacją cech mówcy na podstawie cech głosu. Przedstawiono zarówno systemy opisu cech zawartych w werbalnej, jak i niewerbalnej części wypowiedzi.
4. Wyłoniono najczęściej używane w badaniach CPJ cechy zachowań, wypowiedzi i parametry akustyczne sygnału mowy (aneksy)
5. Sformułowano ogólne postulaty dotyczące własnego systemu opisu cech parajęzykowych

1. Cechy parajęzykowe – sposoby rozumienia i definiowania

Wyrazem wagi, jaką środowisko naukowe przywiązuje do badań nad cechami parajęzykowymi (CPJ) jest m.in. towarzyszący w ostatnich latach konferencji InterSpeech „Paralinguistic Challenge". Ogłaszając go na rok 2010, Schuller stwierdza, że:

> Most paralinguistic analysis tasks resemble each other not only by means of processing and ever-present data sparseness, **but by lacking agreed-upon evaluation procedures and comparability**, in contrast to more traditional disciplines in speech analysis. (podkreślenie MK)

Stwierdzenie to trafnie podsumowuje zebrane niżej próby definiowania CPJ, tworzenia ich inwentarzy oraz instrukcji dotyczących ich tagowania w korpusach.

Longe [1981:103] zauważa, że nadawca komunikatu realizuje trzy podstawowe funkcje:

> In any communicative situation, the writer/speaker (addresser) has to ensure that the means of communication available make it possible to perform the following functions (Longe 1981:103):
> (a) the **modal function** serving to indicate the writer's attitude toward what is being said and toward the person to whom it is said;
> (b) the **metalinguistic function** serving to define exactly what the writer intends terms to mean or what the terms refer to;
> (c) the **contact function** serving to ensure that the channel of contact or communication remains open.

Oczywiście, można doszukać się we wcześniejszej i późniejszej literaturze mniej lub bardziej różniących się inwentarzy funkcji wypowiedzi lub funkcji językowych (włączając w to klasyczną definicję Jakobsona). Już na tym poziomie pojawia się problem składników wypowiedzi i ich „przynależności" do języka (systemu językowego).

Pierwsze użycie terminów "paralanguage" i "paralinguistic" przypisuje się **Tragerowi [1958].**

1 Proponuje się przyjęcie określenia „parajęzykowe", gdyż termin-kalka „paralingwistyczne" oznaczałby „parajęzykoznawcze", a nie dotyczące para-językowego składnika wypowiedzi).

Powołując się na pracę Cowie'ego [2001], Malika Meghjani [2009] definiuje jako „parajęzykowe" to, co nie jest regulowane normami języka. Dosyć wąsko określa parajęzykowe, akustyczne parametry realizacji wypowiedzi (wysokość tonu i intensywność). Zwraca jednak uwagę na to, że na gruncie identyfikacji emocji konieczne jest łączenie w analizach informacji językowej i parajęzykowej.

> The research for audio-based emotion recognition mostly focuses on two measurements: linguistic and paralinguistic [...]. Linguistic measurement for emotion recognition conforms to the rules of the language whereas paralinguistic measurement is meta-data which are related to the way in which the words are spoken, i.e. in terms of variations in pitch and intensity of the audio signal that are independent of the identity of the words in the speech. The decision regarding the relative utility of these two categories of features for emotion recognition is inconclusive in the literature [...]. Hence, in order to obtain an optimal feature set, researchers [...] have combined acoustic features with language information using a Neural Network architecture.

Lamel i Gauvin [1993] następująco

> By the term non-linguistic we refer to **speech features that give little information about the linguistic content of the message**, such as the gender or identity of the speaker, the accent or even the language spoken. (Knowing the identity of the language is of little help in understanding what is being said if one does not already know the language.) While speaker and language identification have been the subject of long term research, they have typically
> been seen as independent research areas, presenting problems different from those of speech recognition. (podkreślenie MK)

Davies i Widdowson [1974] zrównują komponent paralingwistyczny z komponentem niewerbalnym w komunikacji twarzą w twarz. Z perspektywy ogólnych badań nad komunikacją wydaje się oczywiste, że kanał wizualny dostarcza szeregu wskazówek, które mogą zadecydować o sposobie interpretacji wypowiedzi, stanowiąc jej integralną (gdyż niezbędną do właściwego zrozumienia) część.

Abercrombie zauważał, że chociaż zachowanie parajęzykowej jest zachowaniem niewerbalnym, to nie każdy przekaz niewerbalny jest parajęzykowy.

Longe [1999] twierdzi, że:

> **for nonverbal behaviour to qualify as paralinguistic,** it must
> (a) communicate, and
> (b) be part of a conversational interaction'.
> A nervous twitch of the eyelid during conversation is not paralinguistic, although it is a nonverbalbehaviour, so also many personal mannerisms and tics are not paralinguistic. A wolf 'whistle' communicates, but if it does not enter into conversation, it cannot count as a paralinguistic feature. (podkreślenie MK)

Noeth, W. (1990). Handbook of Semiotics. Indiana University Press, Bloomington & Indianapolis.

Roach (za: Crystal 1969):

> While prosodic features are a necessary component of all speech, paralinguistic features may be absent, and allow for more idiosyncrasy in their realisation.

Powstaje jednak wątpliwość, czy rzeczywiście jest możliwa wypowiedź pozbawiona cech parajęzykowych. Nawet jeśli głos “pozbawiony charakteru”, czy możliwie “neutralny”, to na tym właśnie będzie polegała jego specyfika parajęzykowa – tym będzie się wyróżniał od innych, a zarazem będzie to cecha możliwa do interpretacji jako wynik specyficznego stanu psychicznego mówcy. W głosie można doszukiwać się śladów oddziaływania pewnych trwalszych niż chwilowe emocje czynników, jak np. rysów osobowości [Ivanov et al. 2011].

Roach 1998 zauważa, że:

> “Crystal & Quirk (1964) provide the most detailed classification to date of prosodic and paralinguistic features in English, and it is on their classification that the present system of paralinguistic annotations is primarily based.”

Tutaj warto odnotować, że **w 1998 roku klasyfikacja opracowana w roku 1964 jest nadal uznawana za najbardziej kompletną**.

> Roach 1998: Crystal & Quirk m**ake a gradient distinction between prosodic and paralinguistic features in speech**. At the prosodic end of the scale are features such as intonation, which unambiguously exhibit linguistic patterning, and at the other end of the scale are features such as voice quality or clicking noises which do not (1964: 12). However, the authors emphasise the scalar nature of this distinction and do not propose a precise specification of where the paralinguistic gives way to the prosodic: “it would be prejudging the results of future careful research to make a clear-cut list of features undoubtedly playing a role in linguistic patterning and another list of features undoubtedly ‘beyond’ the limits of describable linguistic structure” (1964: 12).

Definicje cech lingwistycznych, paralingwistycznych, pozajęzykowych nie są w literaturze jednoznaczne. Niektóre z nich zebrała Susanne Schötz:

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

Wydaje się, że wyróżnienie kategorii proponowanych w powyższej tabeli jest wysoce umowne, a zasady ich kategoryzacji niejasne. „Pozajęzykowość" wiąże się np. z brakiem intencjonalności lub z cechami mówcy, na które nie ma on większego wpływu (płeć, wiek). Wśród cech parajęzykowych mieszane są cechy samego mócy (również wymieniane wiek i płeć), czynniki środowiskowe, jak i cechy samych wypowiedzi.

Definiowanie i rozumienie CPJ: Podsumowanie

W świetle powyższych, jak i wielu podobnych wypowiedzi (por. spis literatury), można stwierdzić, że:

a) Badacze znacznie różnią się rozumieniem i sposobem definiowania CPJ (jak również rozumieniem tego, co „pozajęzykowe" i „językowe").
b) Różnice te wynikają często z dosyć fundamentalnych różnic w pojmowaniu języka i komunikacji, ale czasami mają charakter techniczny;
c) Definicje CPJ różnią się szczegółowością i aplikacyjnością.

Proponuje się jednoznacznie przyjąć, że **mówiąc o CPJ w ramach niniejszego projektu będzie mówiło się o cechach samej wypowiedzi**, które w pewien sposób mogą stanowić wyraz **określonego stanu mówcy** (stanu psychofizjologicznego – trwałego lub chwilowego). Precyzyjnie, za „potencjalne" CPJ będzie uznawało się parametry wypowiedzi w dowolnym jej wymiarze, które nie wynikają wprost z „regulacji systemu językowego" i niosą dodatkowe znaczenie, szczególnie zaś informację o stanie mówcy. Termin „parametry" należy rozumieć szeroko, włącznie z obecnością pewnych zjawisk lub ich brakiem (np. elizje, upodobnienia, epentezy, substytucje i inne zjawiska z poziomu fonetyczno-fonologicznego, leksykalnego i składniowego).

Na przykład, „barwa głosu" sama w sobie nie jest cechą parajęzykową – za CPJ uznajemy określony układ wartości parametrów związanych z barwą głosu. Podobnie, intonacja może nieść w sobie informację parajęzykową, ale tylko określony zestaw jej parametrów (określone cechy przebiegu fo) można uznać za „cechę parajęzykową". Sama intonacja, barwa głosu, itd. są obecne w wypowiedzi niemal zawsze (wyjątki – np. szept). Jednak należy uznać, że w warunkach naturalnych nie pojawiają się wypowiedzi bez cech „parajęzykowych".

Należy pogodzić się z istnieniem gradacji „parajęzykowości". Na przykład, intonację sygnalizującą pytajność zaklasyfikowano by zapewne jako językową; jednak jaka konkretnie zmiana jest wywołana czynnikiem systemowym (językowym), a jaki jest udział czynnika parajęzykowego (np, indywidualnego stylu). Gussenhoven [2004], opowiadający się za konsekwentnym rozgraniczeniem intonacji językowej i pozajęzykowej nie daje jasnej odpowiedzi.

Kreiman i współpracownicy (Perception of Voice Quality in HSP) zauważają na przykład, że:

> Laver referred to voice quality as "a cumulative abstraction over a period of time of a speaker characterizing quality, which is gathered from the momentary and spasmodic fluctuations of short-term articulations used by the speaker for linguistic and paralinguistic communication" (1980, p. 1).

2. Dostępne inwentarze oraz systemy opisu cech parajęzykowych w sygnale mowy

Poniżej przedstawiono zarys wybranych inwentarzy, kategoryzacji i/lub opisów (tagowania) CPJ. Należy podkreślić, że autorzy tylko w nielicznych przypadkach ujawniają pełną specyfikację systemów. Dlatego szczegółowość opisu danego systemu zależy w dużej mierze od dostępności

materiałów.

(0) Crystal & Quirk (za Roach 1998) wyróżniają dwie kategorie cech parajęzykowych:

- **Voice qualities** are due to different modes of phonation, and are: normal voice, falsetto, whisper, creak, huskiness, and breathiness.
- **Voice qualifications** are non-linguistic vocal effects running through or interrupting speech and include: laughing, giggling, tremulousness, sobbing, and crying.

(1) The **Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies**

UUSDDfPIS to publicznie dostępny korpus. Twórcy następująco definiują jego przeznaczenie:

> [it is] intended for use in understanding the usage, structure and effect of paralinguistic information in expressive Japanese conversational speech [Mori, Kasua, Nakamura 2010]

Dla opisu stanu emocjonalnego określono (idąc za C*ircumplex Model* Russela (1980)) następujące wymiary stanów emocjonalnych:

1. pleasant-unpleasant
2. aroused-sleepy
3. dominant-submissive
4. credible-doubtful
5. interested-indifferent
6. positive-negative

Przykładowa anotacja stanu emocjonalnego (xml):

```
<EmotionalState>
<Rating AnnotatorID="#23" Pleasantness="7"
Arousal="7"
Dominance="3"
Credibility="7"
Interest="7"
Positivity="7"/>
<Rating AnnotatorID="#26" Pleasantness="5"
Arousal="6"
Dominance="6"
Credibility="5"
Interest="6"
Positivity="6"/>
<Rating AnnotatorID="#27" Pleasantness="7"
Arousal="7"
Dominance="6"
Credibility="7"
Interest="6"
Positivity="6"/>
</EmotionalState>
```

Łatwo zauważyć, że system ten to w zasadzie system tagowania stanów emocjonalnych oraz atentywnych, a nie bezpośrednio CPJ. Opis stanów emocjonalnych na podstawie subiektywnej oceny samych wypowiedzi bez dostępu do mówcy w trakcie nagrań i możliwości jego

obiektywnej oceny, prowadzi do „metodycznego zapętlenia”: Na przykład, intuicyjnie uznajemy, że dana barwa głosu w danej wypowiedzi świadczy o zdenerwowaniu, więc tagujemy ją jako taką, a następnie w ramach analizy tagów uzyskujemy wynik, iż dana barwa głosu świadczy o zdenerwowaniu (w istocie wynikiem jest to, że zdaniem anotatora dana barwa głosu świadczy o zdenerwowaniu).

(2) **Anotowany korpus dialogów telefonicznych (LUNA)** [red. Marciniak 2010] - informacja o komunikacji miejskiej
Jedynie w opisie zbioru tagów wzmiankuje się jednostki poza- i parajęzykowe:
- odcinek ciszy powyżej 1 sekundy [silence] <Event desc=”silence” type=”noise” extent=”momentaneous”
- “dźwięki parajęzykowe” (w jednej kategorii z zakłóceniami, których źródłem jest osoba mówiąca”, np. mhm, aaa) [lex=filler]

System charakteryzuje bardzo pobieżne i uproszczone potraktowanie informacji para- i pozajęzykowych. Takie podejście nieco zaskakuje, szczególnie w przypadku korpusu tego typu (mowa spontaniczna). Iloczas pauzy jest istotny, istotne percepcyjnie i „znaczące” mogą być pauzy poniżej jednej sekundy. Kategoria „noise” też budzi wątpliwości., Podobnie jak umieszczenie quasi-słowa „mhm” to kategorii „filler”. Niezrozumiałe jest również włączenie takich intencjonalnie realizowanych i znaczących komunikacyjnie jednostek do kategorii „zakłócenia”. Podobnie uproszczone podejście do cech parajęzykowych zaproponowano w specyfikacji SpeeCon (prawdopodobnie tagowanie polskiej części korpusu LUNA w pewnej mierze opiera się na SpeeCon [Marasek, Gubrynowicz 2004]).

(3) **Multimodalny korpus DiaGest2** (zadanie “origami”, rejestracja 2-4 kamerami + niezależna rejestracja fonii, 20 pięciominutowych dialogów opracowanych, ponad 80 zarejestrowanych) [Karpinski et al 2009-2010]. System transkrypcji oparty m.in. na wcześniejszych pracach, m.in. korpusie PoInt i PCNC [Karpinski et al 2008-2009]

W transkrypcji ortograficznej:

[pc{t}], pauza cicha (t - opcjonalnie podawany iloczas w ms, np. [pc220])
[pw{t}], pauza wypełniona (j.w.)

otwarty zbiór tagów (sygnalizowanych nawiasem kwadratowym), które obejmują m.in.
[śmiech]
[westchnienie]
[jęk]
[oddech] (odnosi się do “głośnych” oddechów, potencjalnie słyszalnych dla współrozmówcy)
[ziewnięcie]

Quasi-słowa, których transkrypcja jest skonwencjonalizowana, zapisuje się bez dodatkowych oznaczeń (np. “aha”, “ach”, “och”, ojej”). Quasi-słowa o dyskusyjnej pisowni lub rzadko zapisywane ortograficznie ujmuje się w nawiasy kwadratowe, np. [yhm],

Opis prozodii (tutaj w zasadzie opis odnosi się do realizacji, bez przesądzania o językowym czy pozajęzykowym charakterze zjawiska):
- podział na frazy

- prominencje (w ograniczonym zakresie)

Poza tym, w zależności od koncepcji tego, co para/pozajęzykowe, a co językowe, można wymienić pewną liczbę tagów odnoszących się do:
- kierunku spojrzenia (np. [partner_area], [own_area])
- ruchów głowy i tułowia
- gestów realizowanych rękami (szczegółowy opis z uwzględnieniem zmian kierunku ruchu, faz gestu oraz podziału na frazy gestowe)
(mimika nie była tagowana)

(4) **System tagowania korpusu London-Lund:**
http://khnt.hit.uib.no/icame/manuals/londlund/index.htm

(mało detaliczne tagowanie wybranych cech parajęzykowych)
(5) Cechy sugerowane w Paralinguistic Challenge (możliwe do ekstrakcji metodami siłowymi):
Table 3: Provided feature sets: 38 low-level descriptors with regression coefficients, 21 functionals. Details in the text. A indicates those only used for the TUM AVIC baseline. Abbreviations: DDP: difference of difference of periods, LSP: line spectral pairs, Q/A: quadratic, absolute.

Descriptors	Functionals
PCM loudness^{-}	position^{-} max./min.
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0–7]$^{-}$	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff.$^{-}$ 1/2
F0 by Sub-Harmonic Sum.	lin. regression error Q/A^{-}
F0 Envelope	quartile^{-} 1/2/3
Voicing Probability	quartile range^{-} 2–1/3–2/3–1
Jitter local	percentile 1/99
Jitter DDP	percentile range 99–1
Shimmer local	up-level time^{-} 75/90

…..

Table 1: Low Level Descriptors (LLDs) and functionals used throughout

systematic construction of a large acoustic feature space.

(6) Roach et al. (1998, 2000) – propozycja tagowania cech parajęzykowych:

Pitch features Round brackets
Intensity features Dot, round brackets
Tempo features Colon, round brackets
Modes of phonation Braces
Fluid control reflexes Square brackets
Respiratory reflexes Colon, square brackets

Start	End
high(	high)
low(	low)
wide(	wide)
narrow(	narrow)
loud.(	loud.)
quiet.(	quiet.)
crescendo.(	crescendo.)
diminuendo.(	diminuendo.)
fast:(	fast:)
slow:(	slow:)
accelerating:(	accelerating:)
decelerating:(	decelerating:)
clipped:(	clipped:)
drawled:(	drawled:)
precise:(	precise:)
slurred:(	slurred:)
falsetto{	falsetto}
creak{	creak}
whisper{	whisper}
rough{	rough}
breathy{	breathy}
ingressive{	ingressive}
+nasal{	+nasal}
-nasal{	-nasal}
glottal-attack{}	
clear-throat[	clear-throat]
sniff[	sniff]
gulp[	gulp]
click[]	
breath-in[	breath-in]
breath-ex[	breath-ex]
breath[	breath]
laugh:[	laugh:]
tremulous:[	tremulous:]
cry:[	cry:]
yawn:[	yawn:]

(7) **BNC (opis kodowania samplera BNC: Lou Burnard, e.g.** http://www.natcorp.ox.ac.uk/corpus/sampler/index.htm)

BNC, uznawany często za pierwowzór współczesnych korpusów językowych. W jego części zawierającej język mówiony zastosowano dosyć złożony i obszerny system tagów dotyczących (w różnej mierze) CPJ.

Voice quality codes:

- crying (18)
- laughing (1151)
- mimicking (46)
- mimicking refined accent (1)
- reading (327)
- screaming (12)
- shouting (165)
- sighing (31)
- singing (157)
- spelling (24)
- whingeing (3)
- whining (4)
- whispering (139)
- yawning (45)

Poniższa lista wartości tagów „paralinguistics" ilustruje, **jakie rodzaje „zdarzeń" uznawane są „parajęzykowe" i jak szczegółowa jest ich kategoryzacja**:

Paralinguistics:
baby talk (18) belch (18) buzzing sound (1) clapping (82) clears throat (139) clicks tongue (1) cough (451) crying (30) gasp (2) giggle (2) gurgle (7) hiccup (4) howl (1) humming (18) imitates aeroplane (1) imitates banjo (1) imitates bringing up phlegm (4) imitates cat licking (1) imitates clearing throat (1) imitates vomiting (3) imitating engine revving (1) kiss (5) kissing (1) kissing sound (2) laugh (3438) laughing (1) licking sound (1) mimicking cat spitting (1) mimicking gorilla noises (2) mimicking microphone noises (1) mimicking shaving noise (1) noise for paws (1) panting (1) purring noises (1) raspberry (7) scream (14) sigh (78) singing (7) sneeze (27) sniff (35) sound effect (8) sound effects (1) sound of biting (1) spitting sound (1) squeak (1) sucking noises (2) sucking then purring noises (1) tt (1) tut (27) whine (1) whistling (44) yawn (22) yelping sound (1)

Poniżej podano listę wartości, które przybierała zmienna <event description>. Jest to istotne ze względu na możliwość ustalenia jakie zdarzenia „kontekstowe" uwzględniano i w jaki sposób je opisywano – jakie wydarzenia uznawano za istotne i jak szczegółowy był ich opis. **Znajomość kontekstu wypowiedzi pozwala na znaczne podniesienie prawdopodobieństwa trafnego ustalenia „znaczenia" danej CPJ, tj. ustalenia, z jakim aspektem stanu mówcy jest związana.**

Background to following (1) Band (1) Band music (1) Break in enquiry (1) Break in recording (1) Children screaming (1) Dramatic music (1) Dramatic music. (1) End of first tape (2) End of recording (3) End of side (3) Engine noises (3) General chatter (1) Gives her a kiss (1) Gunfire, celebration. (1) Horse racing on the radio (1) Intense gunfire (1) Loud rustling next to microphone (1) Mr Bean record on telly I think (1) Music (3) Music and singing (1) Music and song (1) Pen on paper (3) Plane noises (1) Portuguese speech (31) Reading title of book (1) Rock music (2) Sound of burning and dramatic music (1) Sounds of intense automatic weapons fire. (1) Spanish (1) Tape ends (1) Telephone being dialled, overlaps next part of speech. (1) Theme music leading to triumphant climax (1) Theme music, reprise, to end of job (1) Theme music. Engine noise (1) advert (1) advertisements (7) advertisements and travel news (2) another gap in tape (1) applause (9) baby crying (5) baby screaming (2) baby squealing (1) baby talking (8) background to following (1) banging (2) banging noises (1) barking (2) bell ringing (1) bell rings (2) bibs hooter (1) birds singing/whistling (1) birthdays etcetera (1) blowing kisses (2) blowing nose (1) blows nose (2) boxing on television (1) boys fighting (1) break in recording (6) break in recording while watching video (1) break in tape (5) calling from outside (1) car hooter (1) cat miaows (1) chairs being moved (1) cheering (1) cheering and shouting (1) clapping (24) classroom chatter (11) classroom chatter - barely audible speech (1) clicking computer (4) clicking of computer (1) clock chiming (4) closing music (2) cough (4) dog barking (14) dog barks (4) dog sick noise again (1) dogs barking (2) door opening (1) doorbell (1) doorbell ringing (1) doorbell rings (1) dramatic music (1) drilling noise (1) drums (1) duck noise (1) eating (1) eating dinner (1) end of first side of tape and start of second (1) end of job (1) end of recording (5) end of side (1) end of side of tape (1) end of side one of first tape (1) end of side one of tape (1) end of side one of tape. second side starts part way into tape. (1) end of tape (1) end of tape side two (1) engine noises (1) everyone claps (1) football on television (1) football on television again - changed channels (1) general background chatter continues, but foreground conversation has paused (1) general hubbub as people move forward (1) getting into car (1) going into kitchen (1) hammering something (1) happy children (1) hums tune (1) in another room (1) in background throughout following text (1) in canteen/dining room - very noisy (1) in other room (1) intake of breath (1) interruption for radio commercials (2) introduction music (1) jet passes overhead (1) key sounds (1) knock on door (1) knocking on door (3) laugh (1) laughter (5) loud music (2) loud music playing (1) loud music playing on television (1) loud television (1) machinery noises (2) makes dog whining noise (1) makes growling noise (1) makes noise as if being sick (1) making sound of a plane (1) market place noises (1) microphone hissing and conversation very quiet (1) mimicking (1) mimicking crying (1) mimicking dog barking (1) mumble (1) mumbles (1) murmuring from the floor (1) music (5) music of Waltzing Matilda playing (5) music on in background (1) music on loud (1) music on loud again (1) music playing (1) music playing on television (1) music playing with interruption for radio commercials (1) music very loud (1) news bulletin (2) noise (2) noise like the dog being sick (1) noise of dog biscuits being tipped out (1) now moved to another class? (1) paper crinkling noises (1) phone bleeping (2) phone ringing (1) phone rings (3) piano sound (1) plate stacking (1) playing (1) playing piano (1) playing with baby (1) printer noises (1) printer sounds (2) puking noise. (1) radio on (1) reading computer screen (1) reading from board (1) reading from book/text (1) reading from classroom board (1) reading from hospital form (2) reading from newspaper cutting (1) reading from package (1) reading from receipt (1) reading title of book (1) reads horoscopes (1) recording ends (1) repetitive banging (1) scraping something (1) screaming in background (1) shooting sound on video game (1) shuts door (1) shutting door (1) singing (4) singing along to record (1) singing in background (3) singing to music (2) smacking lips together (1) snooker on the telly (1) sombre music (1) something

falls (1) something smashes (1) sounds of burning. (1) sounds of sporadic fire (1) sounds of violence and gunfire (1) speaking French (1) speaking Spanish (2) spells surname (1) spitting noise (1) sports news (1) students' voices in background (6) sucks teeth (3) talking in kitchen away from microphone (1) talking in other room (1) talking to dog (1) talking to the cat (1) talking with mouth full (3) tape ends (6) tape ends side one and starts side two (1) tape playing (1) tape recording (1) tape stops and starts (2) taps table (1) telephone conversation ends (14) telephone conversation starts (13) telephone ringing (5) telephone rings (1) television comes on (1) television loud (1) television on (5) television on - horse racing (1) television turned over to football commentary (1) television very loud (2) telly on loud (1) theme music to end of recording (1) throws dice (1) tickling little girl (2) too much banging (1) traffic (1) traffic news and advertisements (1) travel news (2) travel news and news bulletin (1) travel news and weather (1) unable to hear conversation for some time (1) very loud television - football (1) video film for 135 seconds (1) video playing (1) walking along corridor - just a lot of chatter (1) watching football on television (1) watching football on television and talking quietly in background (2) weather and travel news (1) whispering to dog (1) with microphone (3) woof (1)

(8) Buckeye Corpus (Ohio State University, Pitt et al 2005)
http://buckeyecorpus.osu.edu/BuckeyeCorpusmanual.pdf
Dostępne opisy zawierają niewiele informacji o tagowaniu zjawisk parajęzykowych - prawdopodobnie są to głównie "spontaniczne" notki tagujących, umieszczane w warstwie "log". Transkrybujący zostają poinstruowani w następujący sposób:

> Hesitation sounds: Use "uh" or "ah" for hesitations consisting of a vowel sound, and "um" "mm" or "hm" for hesitations ending with a nasal sound, depending upon which transcription the actual sound is closest to.
> • Use the spelling "huh" when this is used to mean "you don't say".
> • Yes/no sounds: Use "uh-huh" or "um-hum" (yes) and "huh-uh" or "hum-um" (no) for anything resembling these sounds of assent or denial. Again, the versions with 'm' are used when the speaker's utterance ends with a nasal sound. Use "yeah," "yep," and "nope" if these words are said by the speaker. Use "huh" when it is used to mean "what?".
> Label non-speech sounds with the following special labels:
> <LAUGH> = laughter that is NOT part of any word.
> <LAUGH_word> = laughter that is part of a word, e.g., if a speaker laughs while saying the word help, write <LAUGH_help>.
> <NOISE> = noise not from the speaker, such as microphone pops and background sounds.
> <NOISE_word> = noise not from the speaker that occurs while the speaker is saying a word. Supply the word if you can understand it.
> <VOCNOISE > = vocal noise made by the speaker (clearing the throat, sighs, etc.).
> <VOCNOISE_word> = vocalized noised made by the speaker while saying a word.
> • If a speaker uses and gives meaning to a word that is not an actual word, spell the word out as it sounds.

“Corpus Manual” zawiera stosunkowo obszerną instrukcję, obejmującą opis sposobu tagowania detali fonetycznych (nasalizacja, glottalizacja, palatalizacja, i inne zjawiska, w tym cisza). Tabela podsumowująca transkrypcję “non-speech sounds:

TABLE 4. Non-speech labels

LABEL	DESCRIPTION
<LAUGH>, <LAUGH-word(s)>	laughter, laughter during a word or words
<VOCNOISE>	used for other vocalizations which are not speech
<NOISE>, <NOISE-word(s)>	environmental noise (may occur during a word or words)
<IVER>, <IVER-word(s)>	interviewer speech (may or may not be orthographically transcribed)
<SIL>	pause, non-segmental silence
<EXT-word>	non-fluent lengthening on word
<HES-word>	non-fluent hesitation on word
<CUTOFF>, <CUTOFF-clipping=word>	partially produced word (cutoff word) and/or probable target
<ERROR>, <ERROR-error=word>	word produced with lexical or phonological error and/or probable target
{B_TRANS}	beginning of phonetic transcription
{E_TRANS}	end of phonetic transcription
{B_THIRD_SPKR}	beginning of third talker’s utterance
{E_THIRD_SPKR}	end of third talker’s utterance
<VOICE=modal>	describes the talker’s normal mode of voicing
<VOICE=whisper>	used for whispered speech
<VOICE=creaky>	describes extended creaky or glottalized voicing
<VOICE=breathy>	describes breathy voicing
<VOICE=falsetto>	describes use of a very high pitch register for the talker
<VOICE=imitation>	used when talker seems to be imitating someone else
<VOICE=nasalized>	describes use of nasalization over an extended portion of speech
<IVER_overlap-start>	begin portion where interviewer’s speech overlaps with talker
<IVER_overlap-end>	end portion where interviewer’s speech overlaps with talker
<CONF=L>	phonetic aligner has low confidence in phonetic label(s)
<CONF=M>	phonetic aligner has medium confidence in phonetic label(s)
<SYLSTRESS-word=XXXX>	describes lexical stress for individual syllables in proper names (P = primary, S = secondary, U = unstressed)
<UNKNOWN>	speech is audible but unintelligible
<voiceless-vowel>	vowel is present but voiceless
<EXCLUDE-word_word>	portion of conversation which is not phonetically transcribed (e.g., checking sound levels for recording)
<EXCLUDE-name>	omission of identifying information for a talker

9. Joaquim Llisterri: Transcripción, etiquetado y codificación de corpus orales

http://liceu.uab.es/~joaquim/publicacions/FDS97.html#Transcripcion_y_codificacion

Llisterri proponuje system tagowania oparty na standardach takich, jak EAGLES i TEI:

En la siguiente tabla se resumen los principales elementos propuestos por la TEI para la codificación de corpus orales considerados específicos de este tipo de texto (Sperberg-McQueen y Burnard (Eds.), 1994):

Natomiast w zakresie prozodii, Llisterri rekomenduje system SAMPROSA:

A pesar de la coexistencia de los sistemas anteriormente mencionados en el proyecto SAM, probablemente el conjunto de símbolos más extendido actualmente para la transcripción prosódica sea SAMPROSA (SAM Prosodic Alphabet), propuesto inicialmente por Gibbon (1989) y desarrollado por Wells et al.(1992) hasta llegar a su forma actual, que se presenta en la siguiente tabla, reproducida de Wells (1995).

Elemento codificado	**Marca de codificación en SGML**	**Definición**
Divisiones (*division*)		Unidades intermedias entre el texto y el enunciado que permiten delimitar partes diferenciadas en un texto.
Enunciado (*utterance*)	<u>	Segmento de habla comprendido entre dos pausas o delimitado por un cambio en el turno de palabra; puede incluir además información sobre la superposición (<*overlap*>) de turnos cuando interviene simultáneamente más de un hablante.
Pausa (*pause*)	<pause>	Interrupción de la fonación percibida entre dos enunciados o en el interior de los mismos; puede describirse en términos relativos o indicando su duración.
Vocal (*vocal*)	<vocal>	Elemento vocalizado semi-léxico o no léxico (p.ej. pausas llenas o toses).
Kinésico (*kinesic*)	<kinesic>	Cualquier fenómeno comunicativo no vocal (p. ej. gestos).
Acontecimiento (*event*)	<event>	Cualquier fenómeno identificado en la grabación no necesariamente vocalizado ni con valor comunicativo (p. ej. ruidos de fondo).
Texto escrito (*writing*)	<writing>	Texto escrito que se presenta al hablante durante su intervención.

Cambio (*shift*)	<shift>	Momento en el que se produce un cambio en alguno de los rasgos paralingüísticos - calidad de voz, intensidad, rango tonal, ritmo y velocidad de elocución -; cada uno de los rasgos puede describirse mediante una lista de características.

SAMPROSA	**ASCII**	**Definition**
Local tone		
H	72	High pitch
L	76	Low pitch
T	84	Top pitch (extreme H)
B	66	Bottom pitch (extreme L)
M	77	Mid pitch
+	43	Higher pitch
++	43,43	Much higher pitch
+-	43,45	Peak (upward-downward)
-	45	Lower pitch
--	45,45	Much lower pitch
-+	45,43	Trough (downward-upward)
^	94	Upstep
^^	94,94	Wide upstep
!	33	Downstep
!!	33,33	Wide downstep
= or > or S	61 62 or 83	Level or same tone
Global tone: from Local and Nuclear tone repertoire		
Terminal tone: from Local and Nuclear tone repertoire		
Nuclear tone		
-	45	Level tone (before tone group boundary)
' or / or R	39 47 or 82	Rising tone
` or \ or F	96 92 or 70	Falling tone

`' (etc.)	96,39 (etc.)	Fall-rise
'` (etc.)	39,96 (etc.)	Rise-fall
Length		
:	58	Segment length mark
Stress		
"	34	Primary stress
%	37	Secondary stress
Pause		
...	46,46,46	Silence
Boundary		
$	36	Syllable boundary
#	35	Word boundary
|	124	Tone group boundary (non-directional)
[	91	Tone group boundary (left)
]	93	Tone group boundary (right)
Metasymbols		
-	45	Separator (the underscore, _, ASCII 95, may replace this owing to ambiguity with level tone)
*	42	Conjunctor

http://liceu.uab.es/~joaquim/language_resources/spoken_res/biblio_corpus_orals.html

10. **Anotacja korpusu AMI** [McCowan et al. 2007, Carletta et al. 2007]:
AMI to multimodalny korpus (wizja+fonia) obejmujący rejestrację kilkuosobowych spotkań, realizowanych zgodnie z ustalonymi scenariuszami.

http://corpus.amiproject.org/documentations/transcription/

Rozróżnia się „words" i „vocal noise". Te ostatnie to:

$: laugh
% : cough
# : other prominent vocal noise (including creaky voice, aspiration, yawn, throat clearing, tongue click, etc)

Korpus zawiera również anotację gestykulacji i mimiki. Uchodzi za najwszechstronniej otagowany, publicznie dostępny korpus multimodalny.

2.1. Podsumowanie i wnioski dotyczące budowy własnego systemu

Jednym z podstawowych wniosków jest to, iż – mimo dostrzegania roli komponentu para- i pozajęzykowego – nie poświęca się mu w systemach tagowania dostatecznie wiele miejsca. W zaledwie kilku systemach szczegółowość czy nawet tylko poprawność tagsetów można ocenić pozytywnie. Tym cenniejsze byłoby opracowanie własnego systemu.

Uzasadnione wydaje się wyróżnienie kilku płaszczyzn opisu CPJ. Wyróżnić można informacje zawarte w:

- samym sygnale mowy i możliwe do ekstrahowania bez jego segmentacji (bez odwołania się do jego „językowego" charakteru – barwa głosu, energia, itd.)
- w sygnale mowy, ale możliwe do „odczytania" tylko przez pryzmat systemu językowego, w tym zjawiska z poziomu fonologicznego (tj. w odniesieniu do segmentów „językowych" lub poza- lub parajęzykowych; np. realizacja i lokalizacja akcentu, fokalizacja; parametry realizacji pauz wypełnionych i cichych, etc.)
- w doborze i strukturze słownictwa
- w środkach składniowych

Należy również uwzględnić, że niektóre cechy suprasegmentalne mogą być mierzalne w krótkim oknie czasowym, inne dopiero w relatywnie długim (np. stabilność rytmu, jego zmiany). Barwę głosu można traktować jako parametr chwilowy, ale również analizować jej zmiany (co może przynieść znacznie ciekawsze wyniki).

- zjawiska z poziomu fonetyczno-fonologicznego, związane z realizacją „segmentów fonetycznych" (i brakiem lub zaburzeniami ich realizacji): epenteza, elizja, substytucja, itd.
- zjawiska z poziomu leksykalnego (dobór słownictwa, profil częstościowy słownictwa)

Wydaje się, że w literaturze roli leksyki i składni (które też mogą wiele wnieść w kwestii identyfikacji mówcy) poświęca się mniej miejsca ze względu na konieczność transkrypcji i parsingu – wyniki analiz mogą nie być natychmiastowe. Można jednak przypuszczać, że te elementy są również istotnym składnikiem pozwalającym identyfikować głos/mówcę.

Tagowanie CPJ wiąże się z przyjęciem pewnych strategii segmentacji materiału, szczególnie gdy tagi mają mieć charakter przedziałowy, tj. dotyczyć nie „punktu", lecz „odcinka" wypowiedzi (o niezerowym iloczasie). Wszystko wskazuje, że optymalne systemy segmentacji mogą się poważnie różnić w zależności od poziomu, a segmenty z jednego poziomu nie mieścić się w zakresie tagów z innego. Szczególnie problematyczne jest określanie cech/ pomiar parametrów akustycznych. W zasadzie co do większości parametrów można zadać pytanie, czy ich wartości w krótkich oknach czasowych mogą być interpretowane bez uwzględnienia dłuższego kontekstu, czy istnieje dla danego parametru jakieś minimalne okno czasowe, które można by przełożyć na długość segmentu. Można przypuszczać, że jednoznaczna odpowiedź nie istnieje i że każda

segmentacja powiązana z tagowaniem musi być traktowana indywidualnie.

3. Wybrane ustalenia dotyczące związków między CPJ wypowiedzi a stanami mówcy

W dotychczasowych badaniach ustalono szereg związków między parametrami wypowiedzi mówcy (nie tylko akustycznymi) a stanem psychicznym. Warto w tym kontekście brać pod uwagę następujące uwarunkowania:

- Nie istnieje możliwość jednoznacznego ustalenia stanu psychicznego mówcy na podstawie analizy pojedynczej cechy. Stwierdzeniu typu „mowa osoby w stanie depresji jest spowolniona" można jedynie przypisać wysokie prawdopodobieństwo, a praktycznie przydatne staje się ono dopiero, gdy wiemy, jakie jest normalne tempo mowy tej osoby (zob. 2).
- Przydatne jest odniesienie do „normalnego" stylu wypowiedzi mówcy, aby móc ocenić wartość emocjonalną jego innych wypowiedzi. Jeszcze korzystniejsza jest sytuacja posiadania próbek wypowiedzi mieszczących się w różnych miejscach przestrzeni emocjonalnej – można wtedy stworzyć model przestrzeni „emocjonalno-głosowej" dla konkretnego mówcy poprzez aproksymację lub ekstrapolację przy uwzględnieniu ogólnych tendencji w dużych próbach.
- Jednym z kluczowym problemów jest jednoznaczne ustalenie etykiet (nazw, kategorii) stanów, do których mają się odnosić określone wartości CPJ. Na przykład, istnieje wiele kategoryzacji stanów emocjonalnych, o różnym stopniu szczegółowości i odrębności kategorii. Oczywiście, w gruncie rzeczy problem jest znacznie głębszy niż „nazywanie" stanów, jednak ten jego wymiar wykracza poza zakres niniejszego projektu.

Szczególnie wiele prac dotyczy związków między CPJ a stanami afektywnymi. Tutaj jednak zostaną one wspomniane pobieżnie, gdyż stanowią przedmiot odrębnego raportu. Wiele interesujących badań i spostrzeżeń zawiera nowa książka *Emotion-oriented Systems* [red. Petta, Pelachaud, Cowie 2011]. Ogólną sytuację na polu modelowania emocji oddaje w pewnym stopniu prezentacja Scherera „Theories and models of emotions: A swamp." [2004]

Wolf (1972, za Cabeceran 2008) wylicza następujące własności parametrów sygnału mowy, które mogłbyby być wykorzystywane do identyfikacji mówcy:

- low variability within the same speaker
- good discrimination between speakers

- low variability over time
- independent on health conditions
- not subject to mimicry
- unaffected by environmental and transmission noise
- easily measurable
- occur naturally and frequently in speech

Meghjani [2009] przedstawia następujące zestawienie cech głosu emocjonalnego dla pięciu podstawowych emocji:

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wide	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy	Blaring	Resonant	Irregular	Grumbled

Zestaw LLD (*Low Level Descriptors*) do ustalania wartości emocjonalnej wypowiedzi stosowany przez Schuller et al. :

Low Level Descriptors	Functionals
Pitch (F0)	Mean, Standard Deviation
Energy	Zero-Crossing-Rate
Envelope	Quartile 1/2/3
MFCC Coefficient 1-16	Quartile 1 – Minimum
Harmonics-to-Noise-Ratio HNR	Quartile 2 - Quartile 1
Delta Pitch	Quartile 3 - Quartile 2
Delta Energy	Maximum - Quartile 3
Delta Envelope	Centroid, Skewness, Kurtosis
Delta MFCC Coefficient 1-16	Maximum/Minimum Value/Range
Delta Harmonics-to-Noise-Ratio	Relative Maximum/Minimum Position
	Position of 95% Roll-Off-Point

Johnstone/Scherer [xxxx] – dwa wymiary „postawy” i ich korelaty akustyczne:

Table 1. Estimated marginal means for vocal parameters for which main effects of coping or obstruction were significant.

	coping		obstruction	
	low	high	obstructive	conducive
F0 range (Hz)	38.8	35.4		
Glottal slope			-8.6	-9.2
Heart period (ms)			743	758
Resp. depth	12.0	12.8		
Resp. period (s)	4.0	3.7	4.0	3.8

Devillers i Vasilescu [2003]:

F0 variation (sentence level)					
Labels	*Ang*	*Fea*	*Sat*	*Neu*	*Exc*
range F0 (Hz)	++	+	=	=	-
max δF0 (Hz)	++	=	=	=	-

Table 2: *Trends (Mean Values) for emotion effects on selected prosodic parameters correlated with emotion classes for perceptual test subset (40 speaker turns). Symbols: ++: very high, +: high, =: medium, -: low. Ang= anger/irritation, Fea=fear/anxiety, Sat=satisfaction, Neu=neutral, Exc=excuse.*

F0 variation (sentence level)					
Labels	*Ang*	*Fea*	*Exc*	*Sat*	*Neu*
number of sentences	253	192	51	167	4295
range F0 (Hz)	**220**	**228**	201	174	171
max δF0 (Hz)	**129**	**127**	97	91	81

Table 3: *Mean values for emotion effects on selected prosodic parameters correlated with the 5 emotions on the full corpus (5K speaker turns). Symbols: Ang= anger/irritation, Fea=fear/anxiety, Sat=satisfaction, Neu=neutral, Exc=excuse.*

Burkhardt i współpracownicy [2006] mierzą m.in. związek emocjonalności głosu z wartością jittera:

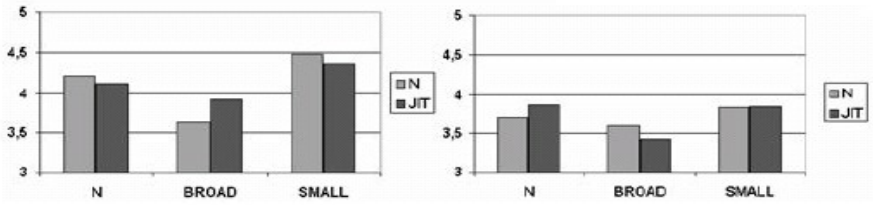


Figure 1: Effects of pitch and jitter for sad (left) and frightened (right) message (N: neutral/original, JIT: jitter)

Schroeder [2001] zestawia najczęściej uwzględniane w syntezie mowy emocjonalnej parametry akustyczne:

Emotion Study Language Rec. Rate	**Parameter settings**
Joy [9] German 81% (1/9)	**F0 mean**: +50% **F0 range**: +100% **Tempo**: +30% **Voice Qu.**: modal or tense; "lip-spreading feature": F1 / F2 +10% **Other**: "wave pitch contour model": main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness [4] American English 91% (1/6)	**F0 mean**: "0", reference line "-1", less final lowering "-5" **F0 range**: "-5", steeper accent shape "+6" **Tempo**: "-10", more fluent pauses "+5", hesitation pauses "+10" **Loudness**: "-5" **Voice Qu.**: breathiness "+10", brilliance "-9" **Other**: stress frequency "+1", precision of articulation "-5"
Anger [6] British English	**F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear [9] German 52% (1/9)	**F0 mean**: "+150%" **F0 range**: "+20%" **Tempo**: "+30%" **Voice Qu.**: falsetto
Surprise [4] American English 44% (1/6)	**F0 mean**: "0", reference line "-8" **F0 range**: "+8", steeply rising contour slope "+10", steeper accent shape "+5" **Tempo**: "+4", less fluent pauses "-5", hesitation pauses "-10" **Loudness**: "+5" **Voice Qu.**: brilliance "-3"
Boredom [10] Dutch 94% (1/7)	**F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C, avoid final patterns 5&A and 12

Table 1. Examples of successful prosody rules for emotion expression in synthetic speech. Recognition rates are presented with chance level for comparison. Sadness and Surprise: Cahn uses parameter scales from –10 to +10, 0 being neutral; Boredom: Mozziconacci indicates intonation patterns according to a Dutch grammar of intonation, see [10] for details.

Mori et al. 2011: zestawienie opisu emocji w kilku korpusach dialogowych:

Table 1
Existing speech corpora for studying spontaneous dialogue.

Corpus	Language	Data collection method	Model of emotion annotation	Public availability
HCRC map task corpus (Anderson et al., 1991)	English	Map task	Nothing	Available
Japanese map task corpus (Horiuchi et al., 1999)	Japanese	Map task	Nothing	Freely available
PASD simulated spoken dialogue corpus	Japanese	Simulated dialogue under various tasks	Nothing	Freely available
ATR-SLDB (Morimoto et al., 1994)	Japanese	Simulated dialogue under the hotel reservation task	Nothing	Available
CallHome (LDC)	English, Arabic, German, Spanish, Japanese, Chinese	Telephone call to family members or close friends	Nothing	Available
Reading/leeds emotion in speech corpus (Greasley et al., 1995)	English	Interviews on radio/TV programs	Category (happiness, sadness, anger, fear, disgust), freely choiced label	Not available
Belfast naturalistic database (Douglas-Cowie et al., 2000)	English	Interviews on TV chat shows	Dimension (activation–evaluation, rated using Feeltrace), Category (40 words)	Not available
JST/CREST ESP corpus (Campbell, 2003)	Japanese	Daily recording of volunteers' natural spoken interactions	Dimension (activation-evaluation, rated using Feeltrace)	Not available
The Vera am Mittag German audio–visual spontaneous speech database (Grimm et al., 2008)	German	Spontaneous and emotional speech recorded from a TV talk show	Dimension (valence, activation, and dominance, using the self assessment manikins)	Freely available
TNO-Gaming corpus (Truong et al., 2008)	Dutch	Talks with friends during playing multiplayer video game with preset events	Dimension (arousal, valence), Category (12 words) Self and observer ratings	Not available
UU database (this paper)	Japanese	Four-frame cartoon sorting task	Dimension (pleasantness, arousal, dominance, credibility, interest, and positivity)	Freely available

3.1. Związki między CPJ wypowiedzi a stanami mówcy: podsumowanie

1. Szczególnie bogata jest literatura dotycząca mowy emocjonalnej. W znacznej części to jednak badania mowy „aktorskiej", przygotowanej lub w sytuacjach laboratoryjnych, bez szczególnej dbałości o „ekologiczność" kontekstu badań.
2. Problemem jest nie tylko identyfikacja samych parametrów akustycznych (i innych) wypowiedzi, które mogą wynikać ze stanu afektywnego, lecz również zbudowanie odpowiedniej taksonomii stanów afaktywnych. Dyskusyjna jest sama kategoryzacja emocji, która może się opierać na różnych modelach (jedno- i wielowymiarowych).Campbell proponuje skupiać się nie tyle na „afekcie" samym w sobie, co na „parametrach nastawienia do rozmówcy" (*attitudinal component*).
3. CPJ nie mające bezpośredniego związku z czynnikiem afektywnym są analizowane znacznie rzadziej. Należą do nich m.in. badania nad wpływem alkoholu, używek, narkotyków, leków, zmęczenia, braku snu, stresu i innego rodzaju obciążeń, patologii układu artykulacyjnego itd. na elementy brzmieniowe wypowiedzi
4. Bada się również ogólniejsze czynniki, które mogą wpływać na sygnał mowy, bez ścisłego przyporządkowania ich do określonych stanów emocjonalnych (np. bardziej napięta lub rozluźniona artykulacja).
5. Związki CPJ (w sensie akustyczno-fonetycznych parametrów wypowiedzi, ale i na innych poziomach) z parametrami mówcy (w tym stanami emocjonalnymi) są w dużej mierze uwarunkowane kulturowo oraz przez sytuację społeczną (Yaeger-Dror 2002 i wiele innych prac). Znaczną rolę odgrywa też element idiolektalny. O

„uniwersalności” pewnych cech można mówić w ograniczonym zakresie, po przyjęciu szeregu założeń.

6. Kolejny fundamentalny (wspominany już wcześniej) problem to kwestia odcinka fundamentu fonetycznego wypowiedzi, na którym należy/można dokonać analizy, czyli „okna czasowego”. Pamiętając, że prócz poziomu fonetyczno-akustycznego warto skorzystać z innych poziomów wypowiedzi (leksykalny, frazowy), można uznać, że minimalną jednostką mogłaby być fraza. Jednak pewne zjawiska (np. zmiany tempa) są wyraźnie obserwowalne dopiero w jeszcze dłuższych oknach czasowych.

W aneksie (C) zawarto listę („mierzalnych”) parametrów akustycznych sygnału mowy, którym najczęściej przypisuje się związek ze stanem psychofizycznym mówcy. Często okazuje się, że stosunkowo „proste” parametry są najlepszymi korelatami cech mówcy (np. iloczas segmentalny i SPL dla wieku, zob. Schoetz i Mueller).

4. Inwentarze oraz systemy opisu cech zachowań niewerbalnych: kanał wizualny

W ostatnich latach często odchodzi się od terminów “kanał niewerbalny” i „komunikacja niewerbalna” na rzecz „kanału wizualnego”. Wynika to m.in. z niemożności precyzyjnego określenia, czym jest „komunikacja werbalna”. Często odnotowuje się również, że „statyczne” opisy np. mimiki, ale i gestu (opis określonej pozycji, ew. sekwencja opisu kolejnych pozycji, ale bez opisu samych przejść) jest niepełny i mylący – podważa się nawet wiele znanych prac, które były oparte na takich metodach opisu. Z drugiej strony, większość dostępnych systemów to właśnie opisy oparte raczej na opisie sekwencji stanów niż samych ruchów; niekiedy istnieje możliwość określenia charakteru ruchu.

W bezpośredniej komunikacji twarzą w twarz z kanału wizualnego najistotniejszą rolę odgrywają następujące parametry:

1. mimika, w tym spojrzenie
    a. ułożenie ust (zwykle trzy wymiary)
    b. ułożenie brwi
    c. otwarcie oczu
    d. kierunek spojrzenia
2. pozycja/ruch głowy
    a. oś przód – tył
    b. oś lewa - prawa
3. pozycja/ruch korpusu
4. pozycja/ruch nóg
5. pozycja/ruch ramion (wiele systemów opisu, od bardzo ogólnych po detaliczne)

Wśród podejść do opisu gestykulacji można wyróżnić:

- podejście "statyczne": stany w określonych momentach (bez opisu przejść)
- podejście "dynamiczne": przejścia (ruchy)
    - "numeryczne" opisy trajektorii ("obiektywne")
    - kategoryzacja (przypisywanie ruchu do wybranej spośród przyjętych kategorii)

Można również mówić o systemach:

- kinetycznych – skupionych na opisie ruchu jako takiego
- funkcjonalnych, semantycznych – skupionych na przypisaniu ruchom pewnych znaczeń

Kolejne możliwe rozgraniczenie podejść to:

- podejście „kwantytatywne", ilościowe: wymagające kategorialnego (np. w skali liczbowej) określania stopnia lub natężenia pewnych zjawisk
- podejście jakościowe: opisuje się jedynie fakt zaistnienia danego zjawiska, bez określania natężenia lub podawania stopniowalnych cech

Poniżej dokonano przeglądu wybranych systemów opisu ruchów ciała i mimiki.

1. **MUMIN** [Allwood et al. 2005]

System skupia się głównie na opisie trzech ogólnych funkcji komunikacyjnych:

- reakcji zwrotnych interlokutora (*feedback: give vs. elicit*),
- zarządzania turami konwersacyjnymi (*turn management*) oraz
- porządkiem komunikacji (*sequencing*)

Modality	**Expression type**
Facial displays	Eyebrows Eyes Gaze Mouth Head
Gestures	Hand gestures (Body posture)
Speech	Segmental (Suprasegmental)

Table 1: Unimodal annotation level

2. **CoGest** [Gut et al 2002].

CoGesT stworzono z myślą o opisie komunikacji multimodalnej, szczególnie zaś dialogu. Wyróżniane kategorie opierają się na kryteriach językowych, a anotacja jest „czytelna" dla człowieka. Podkreśla się odrębność opisu funkcjonalnego i kinetycznego, co daje systemowi cechę uniwersalności (gdyż sama funkcja gestu może być uwarunkowana kulturowo). Gesty są opisywane według następującego schematu:

1. Punkt początkowy
    - lokalizacja
    - kształt dłoni
2. Trajektoria
    - kierunek
    - kształt
    - kształt dłoni
    - modyfikatory
        - wielkość
        - powtórzenia
        - szybkość
3. Punkt docelowy
    - lokalizacja
    - kształt dłoni
4. Symetria

3. ANVIL wraz z systemem tagowania [Kipp, Neff, Albrecht 2007]

System tagowania oraz program (darmowy, podobny w idei działania do ELANa)

http://www.anvil-software.de/

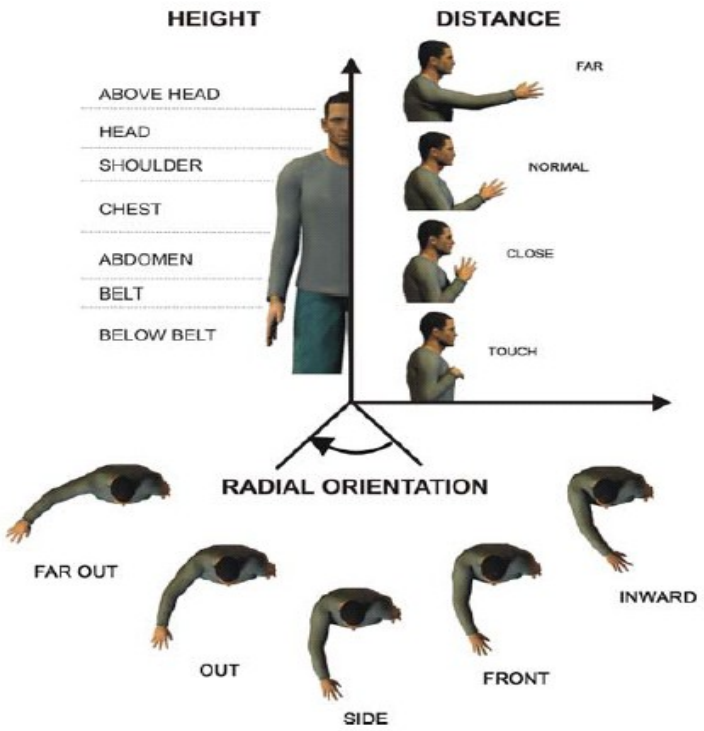


Figure 3. Our three dimensions for hand position.

Wymiary dla określania pozycji ręki. (Kipp M., Neff M., Albrecht I. 2007)

Dodatkowy wymiar opisu – obrót (Kipp, Neff, Albrecht 2007)

4. **Bern System** [Frey 1999]
System o wysokim stopniu szczegółowości, przez co niezmiernie pracochłonny (lecz pozwalający na precyzyjną rekonstrukcję gestów): np. gestu trwającego trzy sekundy system ten koduje siedem punktów czasowych, z których każdy ma dziewięć wymiarów (łącznie 63 atrybuty). Tymczasem bardziej ekonomiczny system Kippa operuje na maksymalnie dwunastu atrybutach.

5. FACES - Ann Kring, Denis Sloan (wersja nieoficjalna - nie można cytować bez pozwolenia autorów)
http://ist-socrates.berkeley.edu/~akring/FACES%20manual.pdf

6. FORM: An Extensible, Kinematically-based Gesture Annotation Scheme [Craig Martell 2002]

Najwyższym poziomem anotacji są *grupy*, obejmujące *podgrupy*. W każdej podgrupie znajdują się *ścieżki. Każ*da ścieżka obejmuje listę *atrybutów* dotyczących konkretnej części ręki lub ciała. Najniższy poziom to poziom konkretnych wartości:

Grupa (Group)
Podgrupa (Subgroup)
Ścieżka (Track)
ATRYBUT (ATTRIBUTE)
Wartość (Value)

Poniższy przykład ilustruje lokalizację atrybutu PODNIESIENIE RAMIENIA z uwzględnieniem grupy i podgrupy do jakiej należy, ścieżki na jakiej się znajduje oraz wartości jakie może przyjąć. Atrybut ten określa kąt między górną częścią ramieniaa ciałem. Wartości podaje się z krokiem rzędu 45 stopni.

Grupa: Prawa/lewa ręka

Podgrupa: Górna część ramienia

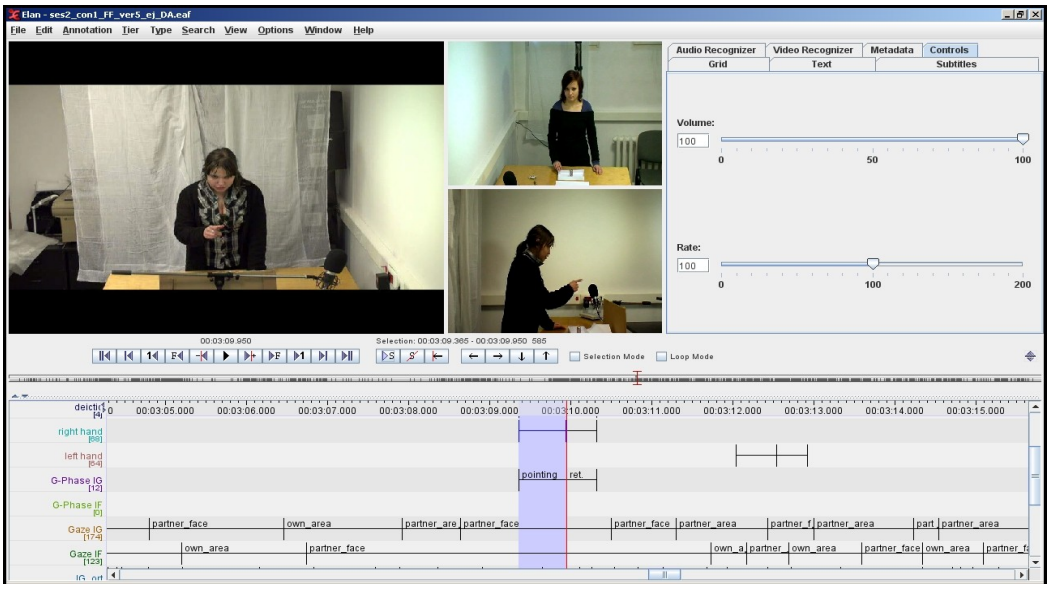
Ścieżka: Lokalizacja

ATRYBUT: PODNIESIENIE RAMIENIA (UPPER ARM LIFT) (z boku ciała)

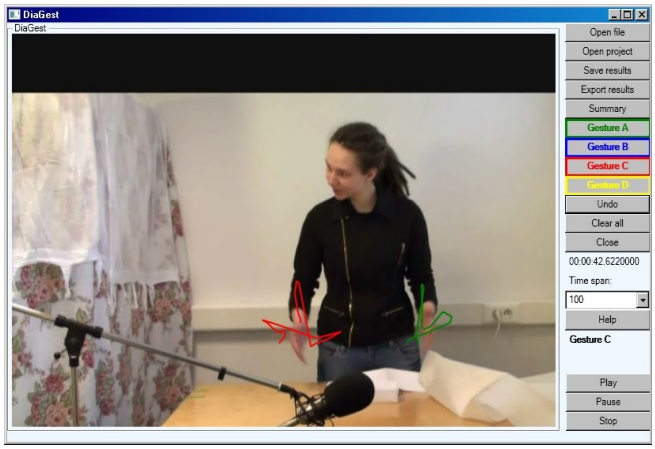
Wartości: no lift (pozycja spoczynku)
0-45 (stopni)
approx. (około) 45
45-90
approx. (około) 90
90-135
approx. (około) 135
135-180
approx. (około) 180

7. DiaGest Coding Scheme [Karpiński et al. 2008]

Wielopoziomowe tagowanie dialogów (stronę gestową opracowała Ewa Jarmołowicz-Nowikow – jest to połączenie tagowania na poziomie frazy gestowej i jej faz z tagowaniem na poziomie ruchu - bez "wstępnej interpretacji" - oznaczaniem punktów zwrotnych kolejnych ruchów). W odrębnych warstwach tagowano, w miarę potrzeby, szczególne kategorie gestów (np. deiktyczne). System był wykorzystywany do tagowania monologów i dialogów w systemie ELAN. Otagowano około 20 dialogów i 40 monologów.



Pomocniczo stosowano prosty program służący do poklatkowego ręcznego znakowania przemieszczenia ramion i dłoni.



8. **Annotation of Human Gesture using 3D Skeleton Controls** [Quan Nguyen, Michael Kipp 2010]

System oparty na rekonstrukcji gestów przy użyciu trójwymiarowego szkieletu, podobnego do służących do animacji. System posiada liczne zalety – jest mniej pracochłonny i – poprzez zdefiniowanie własności szkieletu – pomaga uniknąć wielu błędnych tagów („gestów niemożliwych") i przyspiesza pracę. Podobną koncepcję można zrealizować przy wykorzystanie niemal każdego programu do animacji szkieletowej.

9. FACSAID: Facial Action Coding System Affect Interpretation Dictionary

Specjalizowany system do symbolicznego opisu mimiki (z możliwością dekodowania opisu symbolicznego np. na potrzeby animacji)

http://face-and-emotion.com/dataface/facsaid/description.jsp

10. ANR Multimodality research project 2005-2009: Multimodal Data Transcription and Annotation with ELAN

Autorzy: Jean-Marc Colletta, Olga Capirci, Carla Cristilli, Susan Goldin-Meadow, Michèle Guidetti & Susan Levine
(wymieniona jest też Magdalena Augustyn jako jedna z kilkunastu konsultantek)

11. NEUROGES [Lausberg, Sloetjes 2009].

System obejmujący trzy komponenty opisu (rzadkość!): od czysto kinetycznego po funkcjonalny. Zawiera też procedury opisu – wykracza zatem poza sam zestaw tagów. Procedury tagowania sformułowane są algorytmicznie (przykład poniżej). Technicznie opiera się na wykorzystaniu platformy ELAN (MPI Nijmegen).

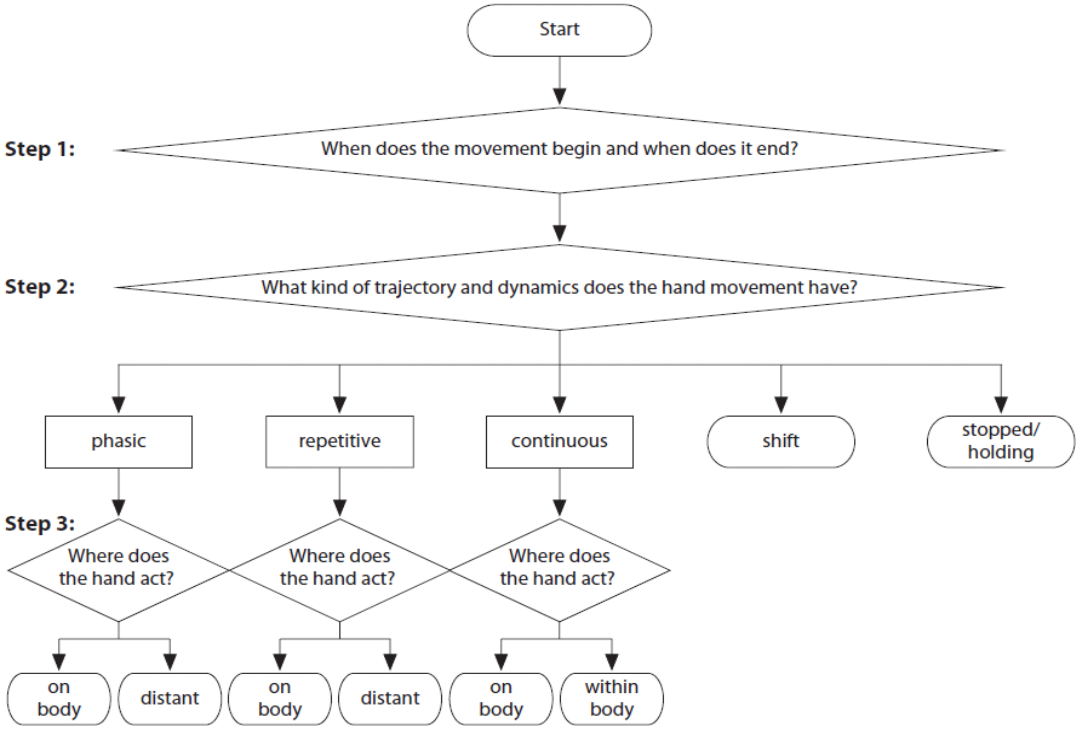


Figure 1. Algorithm for kinetic right-/left-hand coding (Module I).

12. TEI (Text Encoding Initiative) (jako ciekawostka)

Rekomendacje TEI powstawały w okresie, gdy korpusy multimodalne były rzadkością, a zainteresowanie cechami parajęzykowymi niewielkie. Warto jednak odnotować, że istnieją możliwości zapisywania zgodnie z rekomendacjami TEI nawet gestów (kategoria KINESIC: „any communicative phenomenon, not necessarily vocalized").

Przykład:

- <u who="#TVStar">Please,
- <kinesic>
- <desc>raising hand</desc>
- </kinesic>
- let me continue</u>

Podsumowanie systemów tagowania gestu, mimiki i postawy ciała

1. Opisywane tutaj zjawiska są bardzo złożone i jedną z kluczowych kwestii jest ekonomia ich tagowania. Coraz częściej stosuje się tagowanie częściowo zautomatyzowane, oparte na systemach Motion Capture. Systemy takie nie są skuteczne w kategoryzowaniu gestów pod kątem ich postrzegalności, niesionego znaczenia i wartości komunikacyjnej.
2. Systemy znacznie różnią się szczegółowością i liczbą proponowanych kategorii, podejściem do struktury gestu, abstrakcyjnością kategorii, sposobem opisu punktów kluczowych lub trajektorii, wykorzystywanym modelem ruchu.
3. Ze względu na specyfikę, systemy tagowania często łączy się z określonymi programami. Oto lista programów częściej wykorzystywanych w tagowaniu zachowań niewerbalnych: Akira, Anvil, CLAN, ELAN, EXMARaLDA, TASX, MediaTagger, Transana i in. Czasami

stosuje się uniwersalne systemy sprzętowo-programowe, np. **Observer**, obejmujący oprogramowanie, interfejsy wideo, zestawy kamer, itd.

4.1. Ogólne postulaty i wnioski dotyczące budowy własnego multimodalnego systemu opisu CPJ

1. Wielopoziomowa anotacja, obejmująca poziom akustyczny, leksykalny i składniowy, a w ich ramach – odpowiednie podpoziomy (np. w obrębie akustycznego: percepcyjny i „parametryczny"). W razie możliwości/potrzeby, wskazane jest dodanie podpoziomów związanych z mimiką, gestykulacją i postawą ciała. Dzięki wielopoziomowej anotacji i uwzględnieniu elementów językowych staje się możliwe lepsze określenie „wartości" CPJ (np. same pauzy ciche -ich iloczasy i częstość - coś mogą powiedzieć o mówcy i jego stylu mówienia, ale dopiero określenie miejsc w strukturze wypowiedzi i jednostek leksykalnych w okolicach których się pojawiają pozwala powiedzieć coś więcej).
2. Warto zauważyć, że sama informacja płynąca z samego sygnału mowy, jak i z mimiki i gestów rzadko da się zinterpretować ze 100% pewnością bez wiedzy o kontekście.
3. Często nie da się z poziomu akustyczno-pomiarowego ocenić, czy mamy do czynienia z cechą parajęzykową czy językową (o ile warto w ogóle to rozróżnienie utrzymywać). Na przykład, czy dany rosnący kontur intonacyjny ma charakter "językowy" (np. sygnalizuje pytanie) czy też pozajęzykowy (np. sygnalizuje rosnące przerażenie).
4. Podczas gdy wiele cech parajęzykowych można badać lokalnie, np. w obrębie sylaby czy frazy (np. "lokalny pomiar" jakości głosu, wysokości tonu, tempa, etc.), poszerzenie okna czasowego do paratonu lub większych jednostek może pozwolić na zdobycie dodatkowych informacji (zmiany tempa mowy, zaburzenia rytmu, etc.). Generalnie, rozmiar okna czasowego dla danego parametru akustycznego może zależeć od wielu czynników i trudno go z góry zdeterminować. To samo dotyczy
5. Rzadko analizuje się pojedyncze parametry akustyczne. Zwykle są to klastery, liczące od kilkudziesięciu do kilkuset parametrów. To samo dotyczy innych cech – często dopiero ich konfiguracje dają łącznie pożądane wyniki. Z drugiej strony, niekiedy pojedyncze, „proste" parametry okazują się mocnymi korelatami stanu mówcy. Warto więc zachować niezależność opisywania różnych parametrów, a dopiero potem je grupować.

Aneksy do rozdziału I.

ANEKS A

Główne kategorie zjawisk w wypowiedzi mówionej, które można sklasyfikować (całościowo, w pewnym zakresie lub w pewnym kontekście) jako „parajęzykowe”

zjawisko w wypowiedzi (z perspektywy percepcyjnej)	**uszczegółowienie, ewentualnie sposoby „parametryzacji”, miary**	**komentarz**
pauza cicha	liczba pauz cichych na jednostkę czasu liczba pauz cichych w obrębie frazy (średnio) – płynność średnia długość pauzy międzyfrazowej lokalizacja pauz w strukturze składniowej lokalizacja pauz względem jednostek leksykalnych rozkład pauz cichych na przestrzeni dłuższych fragmentów lub całej wypowiedzi	
pauza wypełniona ("jęk namysłu")	realizacja akustyczna (jednorodna, złożona, brzmienie, etc.) liczba pauz wypełnionych na jednostkę czasu średnia długość pauzy wypełnionej lokalizacja pauz w strukturze składniowej lokalizacja pauz względem jednostek leksykalnych rozkład pauz wypełnionych na przestrzeni dłuższych fragmentów lub całej wypowiedzi	
zaburzenia płynności (poza pauzami)	zająknięcia (powtórzenia niepełnych jednostek, np. z... z... zgłosiłem) powtórzenia (w sensie pełnych jednostek, np. wyraz, fraza) zacięcia (szczególny rodzaj pauz, poprzedzonych niedokończoną jednostką)	
oddech	głośność brzmienie oddechu (ciężki, charczący, lekki, świszczący, etc.) częstość loklizacje pauz oddechowych częstość pauz oddechowych	
jednostki leksykalne/paraleksykalne	wulgaryzmy	

	„przerywniki” (jednostki powtarzające się cyklicznie w wypowiedzi) jednostki szczególnie częste (na podstawie rozkładów/list frekwencyjnych) bigramy, trigramy z udziałem jednostek parajęzykowych - w jaki sposób się łączą	chodzi o różnice w stosunku do „profilu odniesienia”
„wokalizacje” para/pozajęzykowe	śmiech westchnienie ziewanie słyszalny oddech jęk (umownie: dźwięczne w zestawieniu z umownie bezdźwięcznym westchnieniem) kaszel chrząkanie (clearing throat) mlaski - "cmokanie" (językowo-podniebieniowe, bilabialne) (lista otwarta)	można wyróżnić odmiany
fonetyczna realizacja frazy	średnia długość frazy (intonacyjnej) średnia długość frazy (mierzona z odstępu między pauzami) zakres zmienności długości frazy intonacyjnej w dostępnym materiale zakres zmienności długości frazy (między pauzami) w dostępnym materiale	
wysokość głosu	średnia wysokość w zarejestrowanym materiale średnia wysokość we frazie zakres zmienności w zarejestrowanym materiale zakres zmienności we frazie średni skok fo w obrębie sylaby akcentowanej stromość przebiegów fo dopasowanie przebiegów do struktury segmentalnej (alignment)	
głośność/energia	zmiana prowadząca do wyraźnych zmian na poziomie artykulacyjnym i brzmieniowym (szept, krzyk) vs. nie prowadząca (głośniej - ciszej)	

	średnia/całkowita energia w obrębie frazy średnia/całkowita energia w obrębie wypowiedzi	
prominencje (akcentuacja)	rozmieszczenie względem „standardowego” - poziom wyrazowy rozmieszczenie względem „standardowego” - poziom frazowy realizacja akustyczna prominencji (składowe: iloczas, energia, zmiany fo) względna energia sylaby akcentowanej	
rytm/tempo	liczba sylab na jednostkę czasu liczba słów na jednostkę czasu liczba prominencji na jednostkę czasu miary równomierności typu (n)PVI (na poziomie segmentów/głosek, sylab, etc.) przyspieszanie/zwalnianie w obrębie frazy przyspieszanie/zwalnianie w obrębie paratonu przyspieszanie/zwalnianie w obrębie całej zarejestrowanej wypowiedzi	
cechy związane ze specyficzną realizacją segmentów	realizacja plozji (w różnych miejscach/sposobach artykulacji) udźwięcznianie i ubezdźwięcznianie + wiele innych możliwości	w zasadzie pełne spektrum możliwych zjawisk fono/fone
tryb fonacji/barwa głosu	ciepły – chłodny szept charczący chrypiący głęboki – płytki dźwięczny ciemny – jasny świszczący + inne "subiektywne kategorie" (np. ciepły, jasny)	parametry akustyczne związane z barwą głosu zawarto w aneksie C
składniowa realizacja frazy	typ porządku składników (np. SVO)	

	dopasowanie składników + parametry związane ze strukturą składniową	
pomyłki językowe	substytucja antycypacja perseweracja zamiana malapropizm freudyzm spooneryzm + inne kategorie	wszystkie dotyczą zarówno wyrazów, jak i większych jednostek

ANEKS B

Elementy najczęściej uwzględniane w opisie gestów i mimiki (lista nie jest wyczerpująca):

mimika	oczy otwarte vs. zamknięte brwi wzniesione vs. ściągnięte (zmarszczone) vs. w neutralnym położeniu nozdrza rozchylone vs. neutralne ułożenie ust (kąciki ust uniesione, opuszczone, neutralne, wargi zaciśnięte, wydęte, przygryzione, żuchwa cofnięta vs. neutralna vs. wysunięta
kierunek spojrzenia	dół vs. góra vs. poziomo (względnie) na rozmówcę vs. obok na obiekt vs. obok na lewo vs. na prawo vs. na wprost
pozycja głowy	uniesiona vs. opuszczona (+ ew. pozycje skrajne) lewo vs. prawo vs. prosto przechylona w lewo vs. przechylona w prawo wychylona do przodu vs. wychylona do tyłu
pozycja korpusu	wyprostowany vs. przygarbiony/pochylony skręcony w lewo vs. skręcony w prawo vs. neutr. brzuch wypięty vs. wciągnięty vs. neutralny pochylony w lewo vs. pochylony w prawo vs. wyprostowany wychylony w lewo vs. wychylony w prawo vs. wyprostowany

	ustawiony bokiem (lewym) vs. na wprost vs. bokiem (prawym)
pozycja ramion	opuszczone wzdłuż ciała założone na piersiach założone do tyłu
pozycja nóg	złączone vs. rozstawione zgięte vs. proste wzajemne usytuowanie stóp podkurczone vs. odwiedzione + inne

Opisu samego ruchu dokonuje się najczęściej poprzez opis punktu wyjściowego, docelowego oraz trajektorii. Można to zrealizować numerycznie lub „opisowo". Można też przyjąć inne podejście, np. opierające się na samych opisach trajektorii. Warto odnieść kształt i lokalizację trajektorii do otoczenia, szczególnie zaś do ciała samego gestykulującego. Opis mimiki powinien uwzględniać „przyzwyczajenia mimiczne", które są często cechą osobniczą, a więc pozwalającą identyfikować mówców.

ANEKS C

Parametry akustyczne łączone z CPJ wypowiedzi. (Każdy z parametrów można w znacznym stopniu uszczegółowić, ale nie leży to w ramach tematyki niniejszego raportu.)

	parametry sygnału lub funkcje na nich bezpośrednio oparte	**przykładowe prace, w których je wykorzystano lub omawiano**
1.	średnie wartości formantów	Kienast, Sendlmeier
2.	średni balans spektralny	Kienast, Sendlmeier
3.	średnia wartość / zmiana fo	niemal wszędzie
4.	średnie tempo / zmiana tempa mowy	niemal wszędzie
5.	średnia energia / zmiany energii sygnału mowy	niemal wszędzie
6.	*shimmer* – wartość lokalna / zmiany	Farrus et al. 2010
7.	*jitter* – wartość lokalna / zmiany	Farrus et al. 2010
8.	*spectral slope, spectral tilt*	Stoll & Doddington 2010
9.	*NHR* i *HNR* (częściowo zależne od jittera i shimmera)	Schuller et al. 2007
10.	*noise burst duration*	Sigmund 2000
11.	długoterminowa zmienność widmowa (LTAS?)	Sigmund 2000
12.	momenty widmowe	Gottsmann, Harwardt 2011

13.	*Haar octave coeficients of residues* (HOCOR)	Zheng 2005
14.	*wavelet octave coeficients of residues* (WOCOR)	Zheng 2005
15.	Mel Frequency Cepstral Coefficients (MFCC)	Schuler 2007

II. Emocje w mowie - współczesne standardy i parametryzacje.

W poniższej części raportu przedstawiono opracowanie dotyczące zagadnień:

- Opracowanie opisu współczesnych standardów dotyczących zasobów przeznaczonych do badań nad prozodią mowy nacechowanej emocjonalnie
- Opracowanie opisu zestawu parametrów akustycznych stosowanych w rozpoznawaniu emocji na podstawie cech głosu

1. Co rozumiemy pod pojęciem „emocja”?

Cechą charakterystyczną emocji jest ich *epizodyczny* charakter – emocja przejawia się jako wyraźna zmiana w funkcjonowaniu organizmu wywołana przez jakiś czynnik zewnętrzny (zachowanie innych osób, zmiana sytuacji w jakiej jednostka się znajduje) lub wewnętrzny (myśli, wspomnienia, doznania) o dużym znaczeniu dla jednostki.

Zazwyczaj uznaje się, że *epizod emocjonalny* składa się z trzech komponentów: fizjologicznego pobudzenia, ekspresji motorycznej oraz subiektywnego uczucia. Czasem uwzględnia się dodatkowo czynnik motywacyjny (np. do jakiego stopnia jednostka jest skłonna podjąć działanie znajdując się w jakimś stanie emocjonalnym) oraz procesy poznawcze biorące udział w ocenie zdarzeń wywołujących emocje i w regulacji zachodzących procesów emocjonalnych (zob. Modele kompozycyjne sek. 1.3).

Emocje niosą informację o tym, które bodźce (zewnętrzne lub wewnętrzne) są dla organizmu istotne, gdyż tylko takie bodźce powodują powstanie emocji. Organizm dokonuje więc oceny docierających bodźców pod względem ich znaczenia, a wynik tej oceny określa zarówno funkcjonalną odpowiedź organizmu (np. przystosowanie się do sytuacji/zdarzenia lub próba opanowania jej) jak i charakter zmian organizmicznych i umysłowych, które będą zachodzić w czasie trwania epizodu emocjonalnego.

Aby ostatecznie zdefiniować co rozumiemy pod pojęciem emocja i jednocześnie odróżnić emocje od innych stanów afektywnych uwzględnimy następujące czynniki (tabela 1):

a) intensywność (intensity),
b) czas trwania (duration),
c) koordynacja/synchronizacja zmian w obrębie komponentów stanu afektywnego (synchronization),

d) stopień w jakim zmiany zachodzące w danym stanie są wywołane lub skupiają się na zdarzeniu/sytuacji, która wywołała dany stan (event focus),
e) stopień w jakim zróżnicowana natura danego stanu jest wynikiem wcześniejszej oceny (appraisal elicitation)
f) szybkość z jaką zmiany organizmiczne i umysłowe zachodzą w danym stanie (rapidity of change)
g) stopień w jakim dany stan afektywny wpływa na zachowanie (behavioral impact)

W takim ujęciu **emocje** możemy zdefiniować jako pewnego rodzaju *epizody* (bo są zjawiskiem krótkotrwałym) w trakcie których *zmiany* wywołane zewnętrznym lub wewnętrznym *bodźcem o dużym znaczeniu* dla organizmu zachodzą *jednocześnie* na *kilku poziomach* (synchronization ++ +): fizjologicznego pobudzenia, ekspresji motorycznej oraz subiektywnego uczucia (i być może również w odniesieniu do gotowości do pojęcia działania i procesów poznawczych). Emocje charakteryzują się największą intensywnością spośród wszystkich stanów afektywnych. Ich natura zależy od wcześniejszej oceny konkretnego zdarzenia/sytuacji lub innego bodźca. Mają bezpośredni wpływ na zachowanie jednostki.

Ponieważ pozostałe stany afektywne (nastrój, postawa, nastawienie oraz cechy osobowości) także niosą informacje istotne z punktu widzenia charakteryzacji i identyfikacji mówcy, zostaną one tutaj krótko scharakteryzowane.

1.1. Charakterystyka pozostałych stanów afektywnych

W przeciwieństwie do emocji **nastrój** wyraża się głównie poprzez zmiany w subiektywnych odczuciach, które zachodzą z podobną szybkością, ale są mniej intensywne i trwają dłużej niż w przypadku epizodu emocjonalnego. W odniesieniu do nastroju trudno mówić o jakimś bezpośrednim zdarzeniu/czynniku, które go wywołują. Zmiany nastroju w miejszym stopniu niż emocje wpływają na zachowanie.

Postawa afektywna w relacjach z innymi osobami ma wiele cech wspólnych z nastrojem (intensywność, czas trwania, synchronizacja – zmiany zachodzą głównie w wymiarze subiektywnych odczuć), ale w większym stopniu niż nastrój (i jednocześnie mniejszym niż emocje) jest ona wynikiem jakiegoś zdarzenia/sytuacji i wpływa bezpośrednio na zachowanie. Zmiany w zakresie subiektywnych odczuć charakteryzujące daną postawę afektywną zachodzą z podobną szybkością jak w przypadku emocji.

Nastawienie (afektywnie zabarwione przekonania, preferencje i skłonności) miewa zróżnicowaną intensywność ale jest stanem trwałym. Nie wiąże się ono ze zmianami na żadnym poziomie ani nie jest wywołane przez żaden bodziec. Nastawienie w minimalnym stopniu wynika z wcześniejszej oceny zdarzenia/sytuacji i taki sam jest jego wpływ na zachowanie jednostki.

Cechy osobowości określają charakter danej osoby oraz skłonności do przejawiania pewnego rodzaju typowych dla niej zachowań. Cechy osobowości charakteryzują się najmniejszą intensywnością i największą stałością spośród wszystkich stanów afektywnych. Podobnie jak nastawienie cechy osobowości nie wiążą się ze zmianami na żadnym poziomie ani nie są wywoływane przez żaden bodziec. W minimalnym stopniu wpływają na zachowanie jednostki, ale wpływ ten nie wynika z wcześniejszej oceny zdarzenia/sytuacji.

BRIEF DEFINITIONS OF AFFECTIVE STATES	INTENSITY	DURATION	SYNCHRONIZATION	EVENT FOCUS	APPRAISAL ELICITATION	RAPIDITY OF CHANGE	BEHAVIORAL IMPACT
Emotion: relatively brief episode of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance (e.g., anger, sadness, joy, fear, shame, pride, elation, desperation)	++ → +++	+	+++	+++	+++	+++	+++
Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant)	+ → ++	++	+	+	+	++	+
Interpersonal stances: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (e.g., distant, cold, warm, supportive, contemptuous)	+ → ++	+ → ++	+	++	+	+++	++
Attitudes: relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons (e.g, liking, loving, hating, valuing, desiring)	0 → ++	++ → +++	0	0	+	0 → +	+
Personality traits: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (e.g., nervous, anxious, reckless, morose, hostile, envious, jealous)	0 → +	+++	0	0	0	0	+

*Symbols indicate the degree to which the features are present, with 0 indicating the lowest (absence) and +++ indicating the highest; arrows indicate hypothetical ranges.

Tabela : Cechy różnicujące różne stany afektywne (za: Scherer *Psychological models of emotion*)

3. Modele emocji

Scherer proponuje podział modeli emocji ze względu na liczbę emocji, które model ma wyjaśnić oraz cechy różnicujące emocje. Przedstawia się on następująco:

3.1. Modele wymiarowe

W modelach tych emocje są klasyfikowane ze względu na *subiektywne odczucie* jednostki, która napotyka na jakiś bodziec i przedstawiane na jednej lub więcej skali.

W modelach **jednowymiarowych** bierze się pod uwagę aktywację/pobudzenie (activation/arousal) lub wartościowość (valency). Pierwsze kryterium pozwala na rozróżnienie pomiędzy emocjami, które wywołują duże pobudzenie i gotowość do działania (np. strach, gniew) a takimi, które wiążą się z postawą pasywną (np. wstyd, smutek), natomiast drugie kryterium umożliwia klasyfikację emocji na skali od negatywnych (złość, gniew) do pozytywnych (radość).

Modele **wielowymiarowe** uwzględniają kilka wymiarów, najczęściej 2 – wartościowość i aktywacja/pobudzenie lub 3 – wartościowość, aktywacja/pobudzenie i wpływ (jednostka przeżywająca emocje może pozostawać rozluźniona lub skupiona/uważna; ten wymiar odróżnia np. smutek od gniewu).

Modele wielowymiarowe mają tę zaletę, że pozwalają na analizę znaczenia zjawisk wykraczających poza emocje. Właściwie każde pojęcie językowe i pozajęzykowe można umieścić w takiej dwu- lub trójwymiarowej przestrzeni aby określić jego znaczenie i cechy dystynktywne.

Choć w modelach wymiarowych bierze się pod uwagę możliwość dalszego rozróżnienia pomiędzy emocjami, podstawowym kryterium pozostaje *wartościowość*, gdzie na jednym końcu skali mamy emocje negatywne, na które reakcją jest unik a na drugim pozytywne – reakcją jest zbliżenie (niektórzy wskazują na istnienie specjalnych obszarów mózgu odpowiedzialnych ze te reakcje).

3.2. Modele dyskretne

Modele te zakładają istnienie pewnej określonej liczby emocji podstawowych (zazwyczaj pomiędzy 9 a 14). Emocje znajdujące się poza tym inwentarzem (których istnienia nie wyklucza się) traktowane są jako wtórne i uznaje się, że powstają one w wyniku interakcji pomiędzy podstawowymi obwodami emocjonalnymi. W modelach dyskretnych główny nacisk kładzie się na ewolucyjne i osobnicze uwarunkowanie

reakcji emocjonalnej.

W **modelach obwodowych** przyjmuje się, że zarówno inwentarz emocji jak i rozróżnienia między nimi określają ewolucyjnie rozwinięte obwody nerwowe. Do podstawowych obwodów emocjonalnych należą: wściekłość, strach, nadzieja/wyczekiwanie i panika. Każdy z nich wiąże się z jasno określoną sekwencją zachowań wywołaną stymulacją nerwową.

Podobnie jak w przypadku modeli obwodowych **modele emocji podstawowych** zakładają istnienie pewnej liczby emocji podstawowych, które stanowią element adaptacyjnych strategii emocjonalnych wykształconych w toku ewolucji. Każda z emocji podstawowych wywoływana jest poprzez ściśle określone warunki i charakteryzuje się określonymi wzorcami reakcji na poziomie fizjologicznym, ekspresji i zachowania.

Modele emocji podstawowych (i dyskretnych wogóle) zostały zapoczątkowane przez Darwina („The expression of emotion in man and the animals", 1872/1998), który dla pewnej liczby emocji wykazał ich sposób „działania", historię ewolucyjną oraz uniwersalny charakter w stosunku do różnych gatunków, stadiów rozwoju osobniczego i kultur.

Kontynuują tę myśl Tomkins zaproponował, żeby podstawowe emocje rozpatrywać jako filogenetycznie stałe programy neuromotoryczne działające w ten sposób, że w odpowiedzi na określny bodziec uruchamiany jest wzorzec reakcji począwszy od poziomu fizjologicznego skończywszy na unerwieniu mięśni, w szczególności mięśni twarzy. W odniesieniu do zmian mimiki twarzy Ekman i Izard postulują istnienie dyskretnych i uniwersalnych wzorców tych zmian dla różnych stanów emocjonalnych.

Dykretne modele emocji znajdują oparcie w *języku*, gdyż niektóre spośród pojęć nazywających emocje występują z dużą częstotliwością (złość, smutek, radość, strach), stąd można je uznać w jakiś sposób za podstawowe.

3.3. Modele zorientowane na znaczenie

Modele leksykalne

W **modelach leksykalnych** punktem wyjścia do badania emocji jest *język*. W centrum zainteresowania znajdują się *słowne opisy subiektywnych uczuć* (a więc podobnie jak w modelach wymiarowych bierze się pod uwagę tylko jeden komponent epizodu emocjonalnego). Uznaje się, że reakcja emocjonalna uwarunkowana jest poprzez *wzorzec*

kulturowy. Podstawą do rozróżnienia stanów emocjonalnych są wspólne dla danej grupy społecznej prototypowe reprezentacje umysłowe.

1.1.1. Modele społecznego konstrukcjonizmu

Modele społecznego konstrukcjonizmu główny nacisk kładą na fakt, że o znaczeniu emocji decyduje społecznie i kulturowo uwarunkowany wzorzec zachowania i oceny. Choć modele te nie wykluczają istnienia psychobiologicznych podstaw reakcji emocjonalnej, to uznają je za drugorzędne w stosunku do sposobu interpretacji przez jednostkę sytuacji/zdarzenia wywołującego emocję (uwarunkowanego kontekstem społeczno-kulturalnym) oraz znaczenia danej reakcji emocjonalnej w interakcji społecznej.

3.4. Modele kompozycyjne

Punktem wyjścia w **modelach kompozycyjnych** emocji jest założenie, że emocje są wynikiem *poznawczej oceny* (która niekoniecznie zachodzi w sposób świadomy lub kontrolowany) *wcześniejszego zdarzenia/sytuacji* oraz że różne wzorce reakcji (w wymiarze pobudzenia fizjologicznego, ekspresji motorycznej, subiektywnego uczucia i skłonności do podjęcia lub nie działania) są wyznaczone poprzez wynik tej oceny. Dokonując subiektywnej oceny jednostka rozpoznaje znaczenie jakie ma dla niej dana sytuacja/zdarzenie oraz swoją zdolność do radzenia sobie w tej sytuacji/w obliczu tego zdarzenia.

Cechą wyróżniającą modele kompozycyjne jest nacisk jaki kładą one na wyjaśnienie mechanizmu wzbudzającego emocje, a jest nim właśnie subiektywna ocena sytuacji/zdarzenia dokonana na podstawie uniwersalnego zestawu kryteriów i podlegająca wpływowi czynników kulturowych i indywidualnych. Wyniki procesu oceny dokonywanej na różnych płaszczyznach (np. czy warunki sprzyjają/nie sprzyjają osiągnięciu celu, krytyczność sytuacji, możliwość radzenia sobie w obliczu danej sytuacji/zdarzenia) można zilustrować w wymiarze wartościowości, aktywacji/pobudzenia oraz wpływu (zob. modele wielowymiarowe). Fakt, że w modelach kompozycyjnych zwraca się uwagę na wpływ czynników kulturowych na wynik oceny oraz znaczenie norm i kontekstu społecznego na kontrolowanie emocji przybliża je do modeli konstruktywizmu społecznego.

Niektóre modele kompozycyjne (np. Lazarusa) zakładają istnienie inwentarza emocji

podstawowych (podobnie jak w modelach dyskretnych), gdyż w przypadku pewnych sytuacji/zdarzeń niezależnie od kontekstu społeczno-kulturowego można zaobserwować te same wzorce oceny i w związku z tym te same reakcje emocjonalne.

Z kolei **Scherer** zakłada, że istnieje tyle stanów emocjonalnych ile możliwości różnych kombinacji wyników subiektywnej oceny bodźca, choć nie wyklucza istnienia emocji, które są nadrzędne w stosunku do innych. Stąd też proponuje wprowadzenia pojęcia *emocji modalnych*, będących odpowiedzią na często spotykane wzorce oceny sytuacji/zdarzeń o charakterze uniwersalnym, które dotykają nie tylko człowieka ale również inne organizmy, np. wszystkie organizmy w każdym stadium rozwoju osobniczego napotykają w jakimś momencie na przeszkodę w osiągnięciu celu lub zaspokojeniu jakiejś potrzeby, można więc przyjąć, że frustracja jest stanem emocjonalnym o charakterze uniwersalnym. Podobny charakter mają też dwa główne wzorce reakcji – walka i ucieczka, które są odpowiedzią na złość i strach będące wynikiem oceny poznawczej.

W ramach modeli kompozycyjnych emocje przedstawia się jako zjawiska złożone z kilku elementów i odpowiadających im funkcji, które są powiązane z konkretnymi podsystemami organizmu (Scherer, *Vocal affect signalling*, str. 199).

- podsystem *przetwarzania informacji* (->składnik poznawczy jego funkcją jest ocena bodźca)
- podsystem *wspierania* (support subsystem, ->neurophysiological component, system regulation)
- podsystem *wykonawczy* (->motivational component, podejmowanie decyzji, planowanie i przygotowanie do działania)
- podsystem *działania* (->expressive component, regulacja nerwowo-mięśniowych stanów i procesów, odpowiedzialny za ekspresję motoryczną i działanie)
- podsystem *nadzoru* (->subjective feeling component, reflection and monitoring)

MODELS	MAJOR FOCUS	ELICITATION MECHANISM	DIFFERENTIATION MECHANISM
Dimensional	Subjective feeling	Rarely directly addressed; rudimentary approach–avoidance definition	Degree of similarity on feeling dimensions such as valence and activation
Discrete emotion	Motor expression or adaptive behavior patterns	Rarely directly addressed; typical situations or stimulus configurations	Phylogenetically continuous neuroanatomical circuits or motor programs
Meaning	Verbal descriptions of subjective feelings	Rarely directly addressed; cultural interpretation patterns	Socially shared, prototypical mental representations
Componential	Link between emotion-antecedent evaluation and differentiated reaction patterns	Appraisal mechanism based on a universally valid set of criteria, influenced by cultural and individual differences	Adaptive reactions in motor expression; physiological responses to appraisal results and the action tendencies generated by the results

Tabela : Porównanie współczesnych modeli emocji (za: Scherer *Psychological models of emotion*)

3.5. Jaki model i jaki inwentarz emocji przyjąć w kontekście charakteryzacji/identyfikacji mówcy?

Każdy z modeli opisanych w poprzedniej sekcji ujmuje i wyjaśnia jakiś aspekt rzeczywistości. Wybór modelu, w oparciu o który emocje będą analizowane jest istotnym punktem, gdyż będzie miał on wpływ na charakter teoretycznego prognozowania wzorców ekspresji głosowej (zmian głosu obserwowanych w kontekście konkretnego epizodu emocjonalnego).

W ramach modeli dyskretnych (Ekman) zakładających istnienie emocji podstawowych (które składają się z określonej liczby względnie stałych wzorców neuromotorycznych) nie wykazano do tej pory istnienia powiązania między konkretnym stanem emocjonalnym a ekspresją głosową (choć takie powiązanie wykazano w odniesieniu do ekspresji twarzy).

W oparciu o modele wymiarowe (np. Wundt) badanie wpływu emocji na ekspresję głosową ograniczone jest do wymiaru wartościowości (stany emocjonalne negatywne vs. pozytywne) i pobudzenia (aktywny vs. pasywny) zarówno w dziedzinie produkcji jak i percepcji mowy emocjonalnej.

Modele kompozycyjne, np. **model procesów składowych** (Component Process Model,

Scherer) nie wykluczają istnienia elementu subiektywnego odczucia (centralnego dla modeli wymiarowych) ani emocji, które z pewnych powodów są nadrzędne w stosunku do innych (definiuje się je jako emocje modalne). Jednocześnie jednak modele kompozycyjne uwzględniają inne elementy epizodu emocjonalnego (jak pobudzenie fizjologiczne, ekspresję motoryczną, procesy kognitywne, zob. sekcja Error: Reference source not found) w szczególności zaś wynik oceny poznawczej dokonywanej na kilku płaszczyznach. W ujęciu kompozycyjnym zamiast ograniczać inwentarz emocji do tych *podstawowych* (jak to ma miejsce w modelach dyskretnych) podkreśla się różnorodność stanów emocjonalnych będących rezultatem kombinacji w obrębie możliwych wyników oceny poznawczej oraz modeluje się różnice między emocjami wewnątrz tej samej rodziny.

W przeciwieństwie do modeli dyskretnych i wymiarowych, modele kompozycyjne dają solidne podstawy do rozważań teoretycznych na temat mechanizmów leżących u podłoża *związku pomiędzy emocjami a ekspresją głosową* i pozwalają na wygenerowanie w tym zakresie bardzo konkretnych hipotez, które można zweryfikować empirycznie. Przewidywanie zmian w ekspresji głosowej w oparciu o wynik oceny poznawczej bodźca na kilku płaszczyznach daje podstawy do określenia charakterystycznych dla danej emocji modalnej wzorców zmian w parametrach akustycznych. Z tych praktycznych względów modele kompozycyjne emocji wydają się być najodpowiedniejsze w kontekście charakteryzacji/identyfikacji mówcy.

4. Fizjologiczne zmiany towarzyszące emocjom i ich przełożenie na zmiany w głosie

4.1. Predykcja zmian fizjologicznych w zależności od wyniku oceny bodźca i ich wpływ na cechy ekspresji głosowej

W ramach swojego **modelu procesów składowych** (component process model) Scherer (1986) wyróżnił 5 wymiarów, w których dokonywana jest ocena bodźca (tzw. **stimulus evaluation checks, SEC**) i w zależności od wyniku tej oceny ustalił kierunki zmian w ekspresji głosowej. Scherer zakłada, że każdy podsystem organizmu (zob. sek. Error: Reference source not found, punkt Modele kompozycyjne „skanuje" zewnętrzne i wewnętrzne bodźce, i dokonuje szeregu ocen według ściśle określonych kryteriów. Ocena dokonywana jest kolejno w następujących wymiarach:

- ✓ nowość (novelty),
- ✓ przyjemność (intrinsic pleasantness),

- ✓ znaczenie celu/potrzeby (goal/need significance),
- ✓ zdolność radzenia sobie (coping potential),
- ✓ zgodność z normami/samym sobą (norm/self compatibility).

*4.1.1. Ocena w wymiarze **nowości***

Wynik oceny bodźca w wymiarze **nowości** określa czy we wzorcu zewnętrznego/wewnętrznego pobudzenia zaszła jakaś zmiana, a w szczególności czy dzieje się coś nowego lub czy można oczekiwać, że jakieś nowe zdarzenie (sytuacja) będzie miało miejsce. W przypadku gdy ocena jest pozytywna, to niezależnie od intensywności bodźca dochodzi do przerwania trwających procesów, skupienia uwagi oraz uwrażliwienia mechanizmów czuciowych[2] w celu zebrania informacji dotyczących nowego bodźca lub ustalenia jego znaczenia.

Organizm odpowiada w sposób ukierunkowany (orienting response) na przetwarzanie informacji. Odpowiedź ta wiąże się z pobudzeniem korowym, zwolnieniem akcji serca, zwężeniem naczyń krwionośnych (w wyniku skurczu mięśni gładkich w ścianie naczyń) w narządach peryferyjnych połączonym z rozszerzeniem naczyń krwionośnych głowy, rozszerzeniem źrenic, podwyższeniem stopnia przewodzenia skóry (który jest wyznacznikiem psychologicznego lub fizjologicznego *pobudzenia*), zmianami we wzorcu oddechowym (pobudzenie obszarów układu siatkowego odpowiedzialnych za wdech). Dodatkowo za sprawą somatycznego układu nerwowego układ ciała zostanie zmieniony w taki sposób, aby narządy czuciowe były skierowane w stronę bodźca (uniesione i zwrócone w kierunku bodźca głowa i górna część ciała).

Opisane tutaj procesy nie pozostaną bez wpływu na ekspresję głosową (szczególnie w powiązaniu z odpowiedzią organizmu w związku z oceną bodźca w pozostałych wymiarach). Tak więc w wypadku pozytywnej oceny bodźca w wymiarze nowości można spodziewać się przerwania ekspresji głosowej, jeśli pojawia się on w trakcie jej trwania, aby nie zakłócać oceny bodźca. Pobudzenie obszarów układu siatkowego

2fundamental physical or chemical processes involved in or responsible for conveying a stimulus to the sensory nerve center

może spowodować głęboki, nagły wdech przypominający brzmieniem dźwięk ingresywny, po którym nastąpi zwarcie krtaniowe.

*4.1.2. Ocena w wymiarze **przyjemny/nieprzyjemny***

Ocenę bodźca w wymiarze **przyjemny/nieprzyjemny** można uznać za podstawową a jej wynikiem jest reakcja *zliżenia* albo *uniku* (-> jedzenie) silnie związana z aktywnością ustno-twarzową. Podstawowy mechanizm *uniku* polega na zwężeniu otworu ustnego i nosowego (zaciśnięcie ust, zmarszczenie nosa), natomiast *zbliżenia* – na ich otwarciu. Ponieważ mechanizmy te są podstawowe i występują u wszystkich organizmów można założyć, że są one wrodzone i kontrolowane na bardzo niskim poziomie centralnego systemu nerwowego (CNS).

Ruchy w obszarze ustno-twarzowym, a w szczególności zmiany w układzie gardła (ściśnięcie/rozszerzenie, napięcie/rozluźnienie jego ścian –> podwyższone napięcie krtaniowe) będą znacząco wpływały na ekspresję głosową. W przypadku oceny bodźca jako *nieprzyjemny* dochodzi do ściśnięcia gardła i napięcia jego ścian, a kanał głosowy zostaje skrócony w wyniku cofnięcia i obniżenia kącików ust. W efekcie w sygnale można zaobserwować więcej energii w wyższych pasmach częstotliwości, zmniejszenie szerokości pasm formantów (szczególnie F1) i zmiany w częstotliwościach formantów (F1>, F2 i F3<) oraz nosowość krtaniowo-gardłową. Głos wyprodukowany w takich „warunkach" można określić jako **wąski** (*narrow voice*).

Reakcją na bodziec odebrany jako *przyjemny* będzie rozluźnienie ścian kanału głosowego, rozszerzenie gardła i skrócenie kanału głosowego poprzez cofnięcie i podniesienie kącików ust. W rezultacie obniży się częstotliwość F1, a szerokość jego pasma zwiększy się, będzie można zaobserwować więcej energii w niższych pasmach częstotliwości, może się też pojawić nosowość gardłowo-językowa. Głos wyprodukowany w takich „warunkach" można określić jako **szeroki** (*wide voice*).

*4.1.3. Ocena w wymiarze **celu/potrzeby***

Głównym celem oceny bodźca w wymiarze celu/potrzeby jest:

- ✓ określenie jego znaczenia dla organizmu (czy bodziec jest ważny dla osiągnięcia

celu/zaspokojenia potrzeby – *relevance subcheck*),

- ✓ ustalenie czy bodziec pomoże/przeszkodzi w osiągnięciu celu/zaspokojeniu potrzeby (*conduciveness subcheck*),
- ✓ w jakim stopniu stan wywołany działaniem bodźca odbiega od stanu oczekiwanego (*expectation subcheck*),
- ✓ czy potrzeba/cel jest na tyle pilna/pilny, że wymaga zewnętrznego działania lub wewnętrznego dostosowania się (*urgency subcheck*)

Wynik oceny bodźca pod kątem wymienionych tutaj aspektów będzie decydował o stopniu zaangażowania organizmu oraz o intensywności reakcji ergotroficznych (polegających na ogólnym pobudzeniu współczulnego/sympatycznego systemu nerwowego (SNS), mobilizacją energii w ciele i zwiększeniem zdolności do działania). Pobudzenie ergotroficzne wywołuje szereg procesów spośród których na ekspresję głosową największy wpłw będą miały zwiększenie napięcia mięśni tonicznych (w kontekście ekspresji głosowej będą to mięśnie krtani) oraz zmniejszenie wydzielania gruczołowego (w tym śliny).

Funkcją pobudzenia ergotroficznego jest przygotowanie do działania w obliczu nagłej sytuacji, która ma dla organizmu istotne znaczenie (*relevance subcheck*) i która odbiega od pożądanego przez organizm stanu (*expectation subcheck*), np. atak wroga. Stąd można spodziewać się podwyższonej aktywności oddechowej (oddech szybki i głęboki), której następstwem może być zwiększenie ciśnienia podgłośniowego, co z kolei prowadzi do podwyższenia F0 i zwiększenia amplitudy. Gdy mięśnie krtani są napięte zaburzona zostaje praca wiązadeł głosowych - fonacja rozpoczyna się w sposób nagły, a drgania są nieregularne i aperiodyczne powodując *jitter* (kolejne okresy różnią się między sobą długością), *shimmer* (amplituda kolejnych okresów jest różna) i zwiększenie energii w wysokim paśmie częstotliwości spektrum (500-1000Hz).

Głos produkowany w takich „warunkach" można określić jako **napięty** i (w mniejszym stopniu) **szorstki** (*tense, harsh voice*). Ekspresję głosową charakteryzuje dodatkowo bardzo dokładna i energiczna artykulacja dźwięków, co powoduje przesunięcie masy języka w stosunku do pozycji neutralnej (dla artykulacji danej głoski, szcz. samogłoski) i w efekcie zmniejszenie szerokości pasm formantów (szcz. F1).

Jeśli bodziec zostanie oceniony jako istotny, nagły i odbiegający od pożądanego przez organizm stanu, ale także jako uniemożliwiający zdobycie celu/zaspokojenie potrzeby (*conduciveness subcheck*) to ekspresja głosowa będzie miała cechy **głosu napiętego i wąskiego** (-> bodziec nieprzyjemny). Jeśli bodziec sprzyja zdobyciu celu/zaspokojeniu potrzeby – **głos napięty i szeroki** (-> bodziec przyjemny).

W przypadku gdy wynik zdarzenia (bodziec) zostaje oceniony jako zgodny z oczekiwaniami (*expectation subcheck*) reakcja organizmu będzie bardziej tropotroficzna (choć wciąż z przewagą pobudzenia ergotroficznego), ukierunkowana na rozluźnienie, powrót do stanu równowagi i zachowanie energii. Spośród procesów związanych z taką reakcją na ekspresję głosową największy wpływ będą miały rozluźnienie aparatu głosowego i zwiększone wydzielanie śliny.

Głos wyprodukowany w takich „warunkach" można określić jako **rozluźniony** (*relaxed voice*). Posiada on cechy zarówno głosu napiętego jak i całkowicie rozluźnionego (*lax v.*), jednak z przewagą tych drugich. Cechy głosu: poziom F0 bliski granicy fizjologicznej (raczej niski ton głosu), poziom amplitudy w zakresie od niskiego do umiarkowanego, brak jitter i shimmer, brak wrażenia szorstkości głosu, równomiernie rozłożona energia spektralna (możliwy jedynie niewielki spadek energii w wyższych partiach spektrum), brak zmian w częstotliwościach i szerokościach pasm formantów.

*4.1.4. Ocena w wymiarze **coping potential***

Głównym celem oceny bodźca pod kątem zdolności organizmu do radzenia sobie z nim (coping potential) jest:

- ✓ sprawdzenie kontroli nad bodźcem i jego konsekwencjami (*control subcheck*)
- ✓ sprawdzenie możliwości uwolnienia się spod dominacji bodźca poprzez działanie zewnętrzne (*power subcheck* – postawa walki lub ucieczki)
- ✓ zdolność organizmu do dostosowania się do wyniku działania bodźca poprzez przemianę wewnętrzną (*adjustment subcheck*)

W zależności od wyniku oceny w wymiarze kontroli (*control subcheck*) można się spodziewać przewagi reakcji ergotroficznych lub tropotroficznych.

Pobudzenie ergotroficzne dominuje gdy organizm ocenia, że **ma wpływ** na wynik lub konsekwencje zdarzenia, lub że **może ich uniknąć**, stąd bierze się postawa walki (fight) lub ucieczki (flight). Dodatkowo, pobudzenie ergotroficzne jest tym większe im

mniejsza pewność siebie osobnika i im silniejsze przekonanie, że nie może się on uwolnić spod dominacji bodźca (taki osobnik będzie pozostawał w stanie gotowości i pobudzenia znacznie dłużej niż osobnik pewny swojej siły ->*power subcheck*). Głos, który powstaje w takich warunkach można określić jako **napięty** (*tense voice*) i będzie on miał tę samą charakterystykę akustyczną co w przypadku ekspresji głosowej gdy bodziec jest oceniany jako istotny ale stan wywołany jego działaniem odbiega od stanu pożądanego przez organizm (goal/need significance check -> goal/need relevant and discrepant).

Dodatkowo, gdy wynik *power subcheck* jest pozytywny (->high power) wystąpi wzorzec oddechowy, zachowanie oraz ekspresja głosowa charakterystyczne dla postawy polemicznej (-> złość): głęboki i silny oddech, rejestr piersiowy (chest register phonation), niskie F0 (co z wynika z pracy wiązadeł głosowych, które w czasie fonacji drgają całą swoją masą i na całej długości), wysoka amplituda oraz energia sygnału w całym zakresie częstotliwości. Tego rodzaju ekspresja głosowa została opisana przez Scherer'a jako **pełny głos** (*full voice*).

Negatywny wynik *power subcheck* (->low power) skłania organizm do przyjęcia postawy defensywnej (np. w strachu). Ekspresja głosowa będzie się charakteryzowała: przyspieszonym i szybkim oddechem, rejestrem głowowym (wysokie napięcie mięśni przywodzących, drgają tylko wąskie brzegi strun głosowych, ciśnienie pogłośniowe jest stosunkowo niskie, głośnia pozostaje lekko rozwarta), niską amplitudą i wysokim F0 oraz spektrum z szeroko rozmieszczonymi harmonicznymi i wyższym poziomem energii wyższych składowych harmonicznych. Głos posiadający takie cechy określa się mianem **cienkiego** (*thin voice*).

W przypadku oceny, że wynik działania bodźca pozostaje całkowicie poza strefą wpływu organizmu (*no control,* np. smutek, przygnębienie, rozpacz) dominować będzie reakcja tropotroficzna (ze strony autonomicznego systemu nerwowego) w wyniku której dojdzie do obniżenia napięcia mięśni aparatu głosowego i odechowego (->breathy voice). W rezultacie będzie można zaobserwować następujące cechy akustyczne głosu: niskie F0 i zakres F0, niska amplituda, szum widmowy, nosowość, bardzo niski poziom energii w wysokich pasmach częstotliwości, słaba pulsacja, częstotliwości formantów bliskie wartościom neutralnym (-> niedokładna artykulacja), szerokie pasmo F1. Głos o takich cechach określa się jako **całkowicie rozluźniony** (*lax voice*) i wyraźnie odróżnia od głosu

neutralnego (modal) i rozluźnionego (relaxed).

*4.1.5. Ocena w wymiarze **zgodności z normami/samym sobą***

Ocena bodźca w wymiarze zgodności z normami/samym sobą (norm/self compatibility check) ma na celu określenie czy konkretne działanie pozostaje w zgodzie z:

- ✓ normami społecznymi, kulturalnymi lub oczekiwaniami innych (*external standards subcheck*)
- ✓ wewnętrznymi normami i standardami wynikającymi z koncepcji własnej osoby (*internal standards subcheck*)

Ocena w tym wymiarze jest właściwa tylko ludziom i nie można wskazać żadnych biologicznych mechanizmów, które leżałyby u podstaw związku pomiędzy jej wynikiem a ekspresją głosową. Scherer proponuje wyjaśnienie zmienności cech głosu wywołanej różną oceną bodźca w tym wymiarze w oparciu o kombinację mechanizmów charakterystycznych dla oceny w innych wymiarach. W rezultacie, gdy standardy nie są naruszone ekspresja głosowa ma cechy głosu **szerokiego i pełnego** (*wide, full voice*). Dodatkowo, gdy stan wywołany działaniem bodźca jest zgodny z oczekiwaniami organizmu (*expectation subcheck -> expected*), ekspresja głosowa będzie miała cechy akustyczne charakterystyczne dla głosu **rozluźnionego** (*relaxed voice*), w przeciwnym razie – głosu **napiętego** (*tense voice*).

Gdy jednak okaże się, że standardy zostały naruszone, ekspresja głosowa będzie nosiła cechy głosu **wąskiego i cienkiego** (*narrow, thin voice*). Ocena w wymiarze *coping potential check* dodatkowo może wpłynąć na głos: będzie on **całkowicie rozluźniony** (brak kontroli nad bodźcem/sytuacją -> *lax voice*) lub **napięty** (kontrola -> *tense voice*).

4.2. Aspekty stanów emocjonalnych: wartościowość (valence), aktywacja (activation) i wpływ (power)

Szczegółowe prognozy dotyczące wpływu wyniku oceny w poszczególnych wymiarach (SEC) na ekspresję głosową są trudne do zweryfikowania empirycznie, ponieważ SEC nie występują prawie nigdy w izolacji, a stan emocjonalny organizmu, nazywany za pomocą terminu określającego dyskretną emocję (radość, smutek, strach itd.), jest rezultatem nakładania się na siebie wyników ocen w kolejnych SEC. Tak więc aby określić cechy głosu

charakterystyczne dla danego stanu emocjonalnego należy zbadać jakie wyniki w poszczególnych wymiarach oceny bodźca (SEC) taki stan spowodują. Należy zauważyć, że wyniki ocen w pewnych podwymiarach (subchecks) mają podobny wpływ na ekspresję głosową np. głos **napięty** „spotykany" w sytuacji gdy bodziec jest oceniany jsko istotny i niezgodny ze stanem oczekiwanym przez organizm -> *goal/need significance check, expectation subcheck* oraz dla sytuacji, gdy osobnik pozytywnie ocenia możliwość kontrolowania bodźca lub jego emocji -> *coping potential check, control subcheck*). Stąd też Scherer proponuje utworzenie nowych wymiarów obejmujących takie wyniki: **wartościowości** (*valence*), **aktywacji** (*activation*) i **wpływu** (*power*, zob. sek. Error: Reference source not found, punkt a).

4.2.1. ***Wartościowość***

Odnosi się do kombinacji wyników w wymiarach *przyjemności* i *celu/potrzeby* (->*expectation subcheck*) z uwagi na prawdopodobieństwo uogólnienia się wzorców ekspresji o podłożu biologicznym, które pojawiają się w odpowiedzi na bodziec oceniany jako przyjemny/nieprzyjemny i ich występowanie także w kontekście bodźców, których działanie jest oceniane przez organizm jako zgodne/niezgodne ze stanem pożądanym przez organizm (->expectation subcheck: consistent or discrepant).

Gdy bodziec zostanie oceniony pozytywnie w wymiarze aktywacji ekspresja głosowa będzie miała cechy głosu **szerokiego**. W przeciwnym razie – cechy głosu **wąskiego**, przy czym na początku skali (od wąskiego do szerokiego) można umieścić głos „pojawiający" się w stanie *wstrętu*.

4.2.2. ***Aktywacja***

Odnosi się do kombinacji wyników w podwymiarach *relevance, expectation*, *urgency* (wymiar celu/potrzeby) oraz *control* (wymiar coping potential). Taka kombinacja jest umotywowana faktem, że wynik oceny w każdym z wymienionych tutaj podwymiarów wiąże się z *pobudzeniem ergotroficznym*. Im ważniejszy cel/potrzeba (->relevant), tym bardziej obecny stan jest potrzegany jako odbiegający od stanu pożądanego przez organizm (->discrepant), tym pilniejsza jest potrzeba działania (->urgent) i tym bardziej organizm dostrzega możliwość wpływania na sytuację poprzez własne działanie (->control) – w takiej sytuacji pobudzenie ergotroficzne jest największe (zob. Ocena w wymiarze celu/potrzeby sek. 1.4). Jednakże w przypadku negatywnego wyniku oceny bodźca w podwymiarze *control* (przekonanie o

niemożliwości kontrolowania działania bodźca ani jego konsekwencji) mobilizacja do działania nie jest potrzebna, w związku z tym nad pobudzeniem ergotroficznym będą przeważać reakcje tropotroficzne (zob. Ocena w wymiarze celu/potrzeby sek. 1.4).

Wraz ze wzrostem pobudzenia ergotroficznego wzrastać będzie **napięcie** głosu (-> tense voice), natomiast wraz z jego spadkiem ekspresja głosowa będzie miała więcej cech głosu **rozluźnionego** (relaxed voice). Im większy udział reakcji tropotroficznych, tym więcej w cech charakterystycznych dla głosu **całkowicie rozluźnionego** (lax voice), przy czym na końcu skali (tense-lax) można „umieścić" głos pojawiający się w stanie *smutku*.

O intensywności napięcia/rozluźnienia głosu będzie decydował udział wyników w poszczególnych podwymiarach w całkowitej ocenie bodźca, przy czym najbardziej napiętego głosu można się spodziewać w przypadku oceny *relevant* + *discrepant* + *urgent* + *control* (spadek napięcia na rzecz rozluźnienia, gdy któryś z czynników jest nieobecny).

*4.2.3. **Wpływ***

Odnosi się do wyniku oceny bodźca w podwymiarze *power subcheck* (coping potential). Im silniejsze przekonanie o możliwości uwolnienia się spod dominacji bodźca poprzez działanie zewnętrzne (-> high power) tym więcej cech głosu **pełnego** (full voice) i mniej cech głosu **cienkiego** (thin voice, zob. punkt Ocena w wymiarze coping potential , sekcja 1.4.1).

> Można zauważyć, że wymiary zaproponowane przez Scherer'a pokrywają się z wymiarami zaproponowanymi w ramach modeli wymiarowych (zob. punkt Modele wymiarowe, sek. Error: Reference source not found). Wynika to stąd, że modele te skupiają się m.in. na odpowiedzi organizmu znajdującego się w danym stanie emocjonalnym, której część stanowi ekspresja głosowa (-> wymiar aktywacji). Jednak zredukowanie SEC do trzech wymiarów wartościowości, aktywacji i wpływu nie oznacza, że nie istnieją zróżnicowane wzorce ekspresji głosowej dla emocji dyskretnych, a nie tylko takich, które różnią się wartością oceny w konkretnym wymiarze (np. pozytywny vs. negatywny, aktywny vs. pasywny).
>
> W tabeli 3 przedstawiono hipotezy dotyczące cech głosu w różnych stanach emocjonalnych. Można zaobserwować, że cechy głosu (tj. wąski, delikatnie napięty, cienki) pokrywają się tylko w dwóch przypadkach: niepokoju i wstydu/poczucia winy. Można to wyjaśnić tym, że w obu przypadkach mamy do czynienia ze stanem niepokoju mającym konkretną przyczynę (-> złamanie norm lub własnych zasad) i określone skutki (-> działanie ze strony innych jednostek, konflikt wewnętrzny).

Emotional state	Hedonic valence	Activation	Power
Enjoyment/happiness	Wide	Relaxed	Slightly full
Elation/joy	Wide	Medium-tense	Medium-full
Displeasure/disgust	Very narrow	Slightly tense	Slightly full
Contempt/scorn	Narrow	Slightly tense	Medium-full
Sadness/dejection	Narrow	Lax	Thin
Grief/desperation	Narrow	Tense	Thin
Anxiety/worry	Narrow	Slightly tense	Thin
Fear/terror	Narrow	Extremely tense	Very thin
Irritation/cold anger	Narrow	Medium-tense	Medium-full
Rage/hot anger	Narrow	Very tense	Extremely full
Boredom/indifference	Narrow	Relaxed	Slightly full
Shame/guilt	Narrow	Slightly tense	Thin

Tabela : Typy ekspresji głosowej w różnych stanach emocjonalnych (za *Scherer, Vocal affect expression*).

5. Uwarunkowania ekspresji głosowej emocji przez czynniki push i pull

Ekspresja głosowa we wszystkich stanach afektywnych (włączając emocje) jest kształtowana przez dwa rodzaje czynników: ***push (effects)***, które mają źródło w zmianach fizjologicznych towarzyszących emocjom i (w mniejszym stopniu) innym stanom afektywnym, sterowanych przez autonomiczny i somatyczny układ nerwowy oraz ***pull (effects)***, które są związane ze społeczno-kulturowymi normami dotyczącymi komunikowania afektu, w tym emocji spełniających funkcje adaptacyjne w społecznej komunikacji i interakcji. Można powiedzieć, że ekspresja głosowa (kodowanie) informacji afektywnej jest wynikiem działania dwóch mechanizmów, które można porównać do procesów down-top (push) i top-down (pull).
Biorąc pod uwagę, że procesy fizjologiczne będą prawdopodobnie zależały od wrażliwości jednostki i specyficznego charakteru sytuacji można się spodziewać, że różnice w ekspresji głosowej afektu obserwowane między różnymi mówcami będą wynikiem wpływu czynników push. Odwrotnie będzie w przypadku czynników pull – można spodziewać się małych różnic w ekspresji afektu między mówcami.
Natomiast wzorce ekspresji tych samych stanów afektywnych w różnych kulturach będą różniły się między sobą ze względu na wpływ czynników pull, a podobieństwa będą wynikiem wpływu czynników push.

5.1. Wpływ czynników push (push effects)

Na podstawie zmian fizjologicznych zachodzących w różnych stanach emocjonalnych, Scherer sformułował hipotezy dotyczące zmian akustycznych, które towarzyszą ekspresji głosowej emocji. Zmiany akustyczne wynikają głównie z modyfikacji

a) wzorca oddechowego,

b) napięcia mięśni aparatu fonacyjnego i artykulacyjnego oraz

c) kształtu toru głosowego.

Hipotezy Scherera dotyczą głównie jakości głosu (voice quality), gdyż na podstawie samej tylko fizjologii trudno sformułować takie hipotezy w odniesieniu do intonacji. Co najwyżej można przewidzieć wzrosty i spadki konturu wynikające ze zmian w oddychaniu.

Aby przetestować te hipotezy (zob. Tabela) należy posłużyć się materiałem, w którym ekspresja głosowa emocji będzie ukształtowana wyłącznie (lub przede wszystkim) przez czynniki push.

5.1.3. Animal calls

Odgłosy zwierząt (animal calls) są przykładem ekspresji o charakterze push. Morton badał wpływ czynników push związanych z różnymi stanami emocjonalnymi na wzorce ekspresji głosowej zwierząt i odkrył, że im mniejsze rozmiar i siła zwierzęcia tym wyższe F0 wokalizacji oraz że gdy rośnie poczucie zagrożenia i chęć ucieczki, wzrasta też tonalność dźwięków.

W odniesieniu do ekspresji głosowej emocji u ludzi obserwacje Mortona do można zinterpretować w następujący sposób: poziom F0 jest związany z poziomem uległości/strachu, a szum spektralny (spectral noise) z poziomem złości/agresji. Obserwacje Mortona nie pozwalają na sformułowanie wniosków dotyczących inotnacji poza tym, że w przypadku umiarkowanej złości/agresji można się spodziewać konturów rosnąco-opadających i opadających.

Trzeba jednak zauważyć, że ekspresja głosowa zwierząt jest także kształtowana (choć w ograniczonym i znacznie mniejszym stopniu niż u ludzi) przez czynniki pull: zwierzęta uciekają się do manipulowania głosem w celu wywarcia pożądanego wrażenia (pull effects).

5.1.2. Infant grunts

Poza odgłosami wydawanymi przez zwierzęta, także niektóre z tych wydawanych przez **niemowlęta** (np. **chrząknięcia – infant grunts**) charakteryzują się profilem akustycznym ukształtowanym pod wpływem czynników push z niewielkim lub wręcz żadnym wpływem czynników pull. Dźwięki te są czystym przejawem zmieniających się stanów afektywnych takich jak ból, głód i radość. W przeciwieństwie do nich niemowlęcy płacz, śmiech i gaworzenie w stopniu szczątkowym przejawiają wpływ czynników pull, gdyż mają na celu wywarcie konkretnego

(choć może nie w pełni uświadomionego) efektu.

5.1.3. Affect bursts

Najbliższym odpowiednikiem odgłosów zwierząt i niemowląt w ekspresji głosowej ludzi są **wybuchy afektu**, których cechy akustyczne są wynikiem tylko i wyłącznie pobudzenia fizjologicznego. Ich występowanie jest powszechne, ale ich forma różni się w zależności od wrażliwości konkretnej osoby i natury konkretnej sytuacji. Z uwagi na to, że wybuchy afektu są najbliższym filogenetycznym odpowiednikiem odgłosów zwierząt, ich wzorce akustyczne powinny być porównywalne, włączając kontury F0. Hipotezę tę zweryfikowano pozytywnie w kilku badaniach, których wyniki pokazały, że poziom F0 rośnie wraz ze wzrostem pobudzenia charakterystycznego dla danego stanu afektywnego.

Gdyby przyjąć pewne **kontinuum** wpływu czynników push i pull na ekspresję głosową emocji to na jednym końcu, **push effects**, umieścilibyśmy **wybuchy afektu**, a na drugim, **pull effects – wzorzec afektu (affect emblem)**. Do tych drugich zaliczamy ekspresję głosową i wyraz twarzy, których cechy są zdeterminowane wyłącznie przez społeczno-kulturowe normy, i które w efekcie wykazują duży stopień konwencjonalności.
+dodać o typach baz mowy afektywnej i jak się ma materiał w nich zawarty do badania wpływu czynników typu push

5.2. Wpływ czynników pull (pull effects)

Scherer wyróżnia 4 główne czynniki pull:

5.2.1. Naśladowanie czynników push

Np. ponieważ rozmiar ciała jest skorelowany z wysokością tonu w wyniku ograniczeń anatomicznych (czynnik push), niektóre zwierzęta wykształciły cechy morfologiczne i zachowanie pozwalające im na obniżenie tonu (naśladując w ten sposób inne osobniki). Można się spodziewać, że aktorzy biorący udział w nagraniach baz mowy emocjonalnej/afektywnej będą wykorzystywali tego typu strategię, tj. będą starali się naśladować głos osoby znajdującej się pod wpływem jakiejś emocji lub innego stanu afektywnego. W tym celu mogą polegać na własnych obserwacjach lub wykorzystać metodę Stanisławskiego, a wyprodukowane przez aktorów wypowiedzi emocjonalne będą przejawiały głównie wpływ czynników push. Hipotezy dotyczące cech akustycznych takich wypowiedzi

będą takie same jak w przypadku wybuchów afektu i będą uwzględniały tylko poziom F0 i szum spektralny, a nie intonację.

5.2.2. Społeczne i kulturowe konwencje/normy dotyczące „głosowego przekazywania emocji".

Ich źródła należy szukać w naśladowaniu czynników push, które uległo konwencjonalizacji. Aktorzy „portretujący głosowo" emocje mogą niekiedy ulegać pewnym stereotypom kulturowym w tej dziedzinie, wykształconym np. w szkołach aktorskich i mających niewiele wspólnego z efektami push. Można stwierdzić, że gdy konwencje ekspresji głosowej emocji są podobne w różnych kulturach, to prawdopodobnie naśladują czynniki push, zaś gdy różnią się między sobą oznacza to, że ich cechy zdeterminowane są przez czynniki pull wykształcone przez lokalne społeczne konwencje. W tym wypadku można oczekiwać występowania konturów intonacyjnych typowych dla konkretnych kultur.

5.2.3. Symbolizm dźwięku

Np. nagły skok amplitudy i wysokości tonu dźwięku są kojarzone z agresją, podczas gdy ich łagodne zmiany – ze słabością i rozluźnieniem. Można uznać, że symbolizm dźwięku jest uniwersalny, gdyż określone wzorce akustyczne (np. te tutaj opisane) są pobodnie postrzegane w różnych kulturach. Przykładem jest mowa skierowana do niemowląt, w która charakteryzuje się podobnymi wzorcami akustycznymi dla tych samych komunikatów (zakaz, aprobata, uwaga, pocieszenie itp.) w różnych językach. Jej cechy mogą być do pewnego stopnia wynikiem działania czynników push (opisanych powyżej, także: Ohala frequency code).

5.2.4. Inne kanały ekspresji

Np. jeżeli w danej kulturze grzeczność nakazuje dużo uśmiechu, to można to uznać za czynnik pull, który nie pozostanie bez wpływu na cechy akustyczne ekspresji głosowej, gdyż uśmiech powoduje zmiany w kształcie kanału głosowego i w efekcie zmiany częstotliwości formantów.

5.2.5. Język i intonacja

Język, a dokładnie **wzorce intonacyjne**, które obowiązują w danym języku: intonacja będzie czynnikiem pull, a jej pragmalingwistyczne zastosowanie jest kodowane w sposób konfiguracyjny (np. modele ToBI, IPO) w przeciwieństwie do ciągłego kodowania zmian poziomie i zakresie zmian F0 (czynnik push).

5.2.6. Optymalna transmisja dźwięku

Np. istnieją gatunki, u których związek F0-rozmiar ciała został zerwany w wyniku konieczności

komunikowania się w określonych warunkach (las, sawanna).

6. Parametry akustyczne charakteryzujące ekspresję głosową w różnych stanach emocjonalnych (do analizy i rozpoznawania emocji na podstawie głosu).

Opisane w poprzedniej sekcji zmiany fizjologiczne o istotnym wpływie na ekspresję głosową można analizować poprzez dokonanie pomiarów szeregu parametrów akustycznych. Do podstawowych parametrów należą: *iloczas, amplituda, F0* oraz *rozkład energii w spektrum*. Są one łatwe do zdefiniowania i można je mierzyć automatycznie, a ich wartości wskazują na stopień pobudzenia (-> wynik oceny w wymiarze aktywacji). Aby wyjść poza wymiar aktywacji konieczne jest zastosowanie bardziej złożonych parametrów (tabela 4).

Właściwa interpretacja parametrów akustycznych stosowanych w analizie związku pomiędzy emocjami a ekspresją głosową wymaga odniesienia do procesów fizjologicznych leżących u podstaw wytwarzania sygnału mowy i zachodzących w danym stanie emocjonalnym. Na proces wytwarzania sygnału mowy składają się: oddychanie, fonacja i artykulacja. Ponieważ zmiana położenia artykulatorów działa na zmieniające się w czasie *źródło sygnału* (wytworzone w procesie oddychania i fonacji) niczym *filtr*, którego właściwości również podlegają zmianom w czasie, w opisie cech filtra i źródła sygnału należałoby posługiwać się odmiennymi parametrami. Niestety wyodrębnienie w sygnale tego, co jest wynikiem działania filtra (->artykulacja) od tego, co wynika z charakterystyki źródła sygnału (->oddychanie, fonacja) jest bardzo trudne. Podobnie jest z rozróżnieniem pomiędzy wpływem oddychania a wpływem fonacji na cechy źródła sygnału, dlatego też bardzo często w analizie sygnału mowy oddychanie i fonację traktuje się jako jeden komponent procesu wytwarzania sygnału. W obu przypadkach jest jednak możliwe wyznaczenie (rozłącznych) zbiorów parametrów akustycznych, które będą opisywały właściwości tylko jednego z komponentów procesu wytwarzania sygnału: oddychania, fonacji albo artykulacji.
W tabeli 4 przedstawiono listę takich parametrów akustycznych i przewidywane kierunki zmian w ich wartościach w zależności od stanu emocjonalnego mówcy (w oparciu o wyniki SEC i potwierdzone danymi empirycznymi, zob. *K. Scherer* 1986*, Vocal affect expression*).

- *Parametry źródła głosu* (voice source parameters) podzielone zostały na takie, które opisują *ogólną jakość głosu* (general voice quality) oraz parametry *prozodyczne*. Te pierwsze, w większym lub mniejszym stopniu, odzwierciedlają stopnień napięcia konkretnych mięśni krtani, który wiąże się bezpośrednio z wynikiem oceny bodźca, a więc i ze wzorcem odpowiedzi charakterystycznym dla danego stanu emocjonalnego. Zmiany w napięciu mięśni krtani nie mogą być w żaden sposób kontrolowane przez organizm i pojawiają się niezależnie od tego, czy ekspresja głosowa ma miejsce czy nie. Jeśli tak, to napięcie mięśni krtani nakłada się na ogólne napięcie obecne w innych częściach ciała (->pobudzenie ergotroficzne). Parametry opisujące *ogólną jakość głosu* są są związane z czynnikami *push* (Scherer), które mają podłoże fizjologiczne. Parametry *prozodyczne*, za pomocą których można opisać wzorce prozodyczne charakterystyczne dla wypowiedzi nacechowanych emocjonalnie są związane z czynnikami *pull* zdeterminowanymi społecznie i kulturowo.
- *Parametry artykulacyjne* można podzielić na takie, które są konieczne do wytworzenia zrozumiałego sygnału (np. pozycje formantów) i takie, które odzwierciedlają czynniki pozajęzykowe (np. wyraz twarzy).
- *Tempo i płynność mowy* wyrażone są za pomocą parametrów takich jak: liczba sylab na sek., iloczas sylaby i samogłoski akcentowanej, liczba i czas trwania pauz, względny iloczas segmentów dźwięcznych i bezdźwięcznych.

Analiza parametrów akustycznych wiąże się z koniecznością przeprowadzenia *normalizacji* względem cech danego mówcy, począwszy od wielkości, kształtu i masy narządów aparatu głosowego (->vocal equipment) na stanie zdrowia i indywidualnych zwyczajach mówcy skończywszy (->vocal setting). Trudności w ocenie zmian zachodzących w głosie pod wpływem emocji na podstawie pomiarów akustycznych wiążą się również z tym, że charakterystykę akustyczną zmiennego w czasie sygnału mowy opisuje się za pomocą parametrów *statycznych* takich jak średnia i odchylenie standardowe lub współczynnik zmienności. Aby uchwycić wpływ emocji na cechy ekspresji głosowej za pomocą tego rodzaju parametrów należy dokonywać ich pomiarów na poziomie pojedynczych głosek, a szczególnie samogłosek, gdyż pozwala to zmniejszyć wpływ czynników artykulacyjnych na charakterystykę źródła sygnału. W ekstrakcji cech dynamicznych takich jak F0 lub formanty przydatne byłoby zastosowanie *stylizacji* aby uzyskać uproszczone ale funkcjonalnie identyczne obrazy przebiegów tych cech.

Acoustic Parameters	Arousal/Stress	Happiness/ Elation	Anger/Rage	Sadness	Fear/Panic	Boredom
Speech Rate and Fluency						
Number of syllables per second	>	>=	<>	<	>	<
Syllable duration	<	<=	<>	>	<	>
Duration of accented vowels	>=	>=	>	>=	<	>=
Number and duration of pauses	<	<	<	>	<>	>
Relative duration of voiced segments			>		<>	
Relative duration of unvoiced segments			<		<>	
Voice Source—F0 and Prosody						
F0 mean[3]	>	>	>	<	>	<=
F0: 5th percentile[3]	>	>	=	<=	>	<=
F0 deviation[3]	>	>	>	<	>	<
F0 range[3]	>	>	>	<	<>	<=
Frequency of accented syllables	>	>=	>	<		
Gradient of F0 rising and falling[3,6]	>	>	>	<	<>	<=
F0 final fall: range and gradient[3,4,7]	>	>	>	<	<>	<=
Voice Source—F0 and Prosody						
F0 mean[3]	>	>	>	<	>	<=
F0: 5th percentile[3]	>	>	=	<=	>	<=
F0 deviation[3]	>	>	>	<	>	<
F0 range[3]	>	>	>	<	<>	<=
Frequency of accented syllables	>	>=	>	<		
Gradient of F0 rising and falling[3,6]	>	>	>	<	<>	<=
F0 final fall: range and gradient[3,4,7]	>	>	>	<	<>	<=
Voice Source—Vocal Effort and Type of Phonation						
Intensity (dB) mean[5]	>	>=	>	<=		<=
Intensity (dB) deviation[5]	>	>	>	<		<
Gradient of intensity rising and falling[2]	>	>=	>	<		<=
Relative spectral energy in higher bands[4]	>	>	>	<	<>	<=
Spectral slope[1]	<	<	<	>	<>	>
Laryngealization		=	=	>	>	=
Jitter[3]		>=	>=		>	=
Shimmer[3]		>=	>=		>	=
Harmonics/Noise Ratio[1,3]		>	>	<	<	<=
Articulation—Speed and Precision						
Formants—precision of location	?	=	?	<	<=	<=
Formant bandwidth	<		<	>		>=

Notes:
1. depends on phoneme combinations, articulation precision or tension of the vocal tract
2. depends on prosodic features like accent realization, rhythm, etc.
3. depends on speaker-specific factors like age, gender, health, etc.
4. depends on sentence mode
5. depends on microphone distance and amplification
6. for accented segments
7. for final portion of sentences

In specific phonemes, < "smaller," "lower," "slower," "less," "flatter," or "narrower"; = equal to "neutral"; > "bigger," "higher," "faster," "more," "steeper," or "broader"; <= smaller or equal, >=bigger or equal; <> both smaller and bigger have been reported

Tabela : Parametry akustyczne i przewidywane kierunki zmian w ich wartościach w podstawowych stanach emocjonalnych (za Scherer, *Vocal affect expression*).

Komentarz

Na podstawie wyników przedstawionych w tabeli 4 można stwierdzić, że kierunek zmian parametrów głosowych odzwierciedla stopień pobudzenia (->wymiar aktywacji/pobudzenia).

Ekspresja głosowa w stanach emocjonalnych o wysokim stopniu pobudzenia – szczęście/euforia, złość/wściekłość, strach/panika oraz wysoki poziom stresu charakteryzuje się szybszym tempem wymowy, wyższym F0, większą zmiennością w zakresie intonacji (więcej akcentów, większe spadki i wzrosty F0), wyższą intensywnością i większym zróżnicowaniem w przebiegu zmian intensywności. Kierunek zmian w obrębie tych parametrów jest odwrotny w przypadku emocji o niskim stopniu pobudzenia jak smutek czy znudzenie.

Jednocześnie widzimy, że zmienność w obrębie parametrów artykulacyjnych oraz tych, które opisują typ fonacji (laryngelization, jitter, shimmer, HNR) nie wykazuje związku ze stopniem pobudzenia. Nieregularność w drganiach więzadeł głosowych pojawia się bardzo często w stanie panicznego strachu (płacz, wymioty, „urwanie" głosu), ale nie w złości, choć obie emocje charakteryzuje podobny stopień pobudzenia. Można więc założyć, że niektóre parametry głosowe, np. te opisujące typ fonacji są bezpośrednio związane z typowym dla danej emocji unerwieniem mięśni krtani. Natomiast parametry spektralne odzwierciedlają jakościowe różnice pomiędzy emocjami (-> modele dyskretne).

W odniesieniu do parametrów artykulacyjnych, zmienność w ich obrębie wykazuje raczej związek z konkretnymi emocjami niż z wymiarem pobudzenia (-> cechy aparatu głosowego zmieniają się pod wpływem ekspresji w części twarzowo-ustnej, np. uśmiech w radości, grymas twarzy w złości).

Choć wymiar pobudzenia jest podstawowy w charakterystyce stanów emocjonalnych, nie jest możliwe wyjaśnienie zmian w parametrach głosowych opisujących ekspresję głosową w różnych emocjach w odniesieniu wyłącznie do stopnia pobudzenia (-> emocje o podobnym stopniu pobudzenia, jak np. wściekłość, panika i euforia są z dużą dokładnością identyfikowane zarówno przez słuchaczy jak i automatycznie).

Wyniki badań empirycznych nie dowodzą istnienia typowych dla danych emocji wzorców ekspresji głosowej (któr potwierdziłyby istnienie neuromotorycznych programów dla konkretnych emocji, -> modele dyskretne), ponieważ w obrębie jednej rodziny emocji obserwujemy zróżnicowane wzorce ekspresji głosowej, np. gwałtowna złość (hot anger), ale nie „zimna" złość (cold anger) charakteryzuje się wyższym F0, większą zmiennością w przebiegu F0, szybszym tempem wymowy, wyższym poziomem energii w wysokich pasmach częstotliwości sygnału.

Podsumowując, pomimo licznych badań dotyczących wpływu emocji na ekspresję głosową dotychczas nie zbadano parametrów akustycznych związanych z wymiarami innymi niż pobudzenie-aktywacja, tj. wymiar wartościowości i wpływu-kontroli. Ponadto, choć emocje na podstawie głosu są rozpoznawane z dużą dokładnością zarówno przez słuchaczy jak i przez modele akustyczne, nadal do końca nie wiemy jakie aspekty epizodu emocjonalnego odzwierciedlają parametry akustyczne brane pod uwagę w analizie i rozpoznawaniu emocji[3].

3możemy jedynie przypuszczać, że zróżnicowana ekspresja głosowa emocji jest wynikiem oceny bodźca w wielu

6.1. Badania zorientowane na produkcję

6.1.1. Przegląd ogólny wcześniejszych prac

Już pierwsze badania w omawianym tutaj zakresie wykazały związek pomiędzy wymiarem aktywacji a najczęściej mierzonymi parametrami akustycznymi:średnim F0, średnią intensywnością i tempem mowy. Niektórzy badacze zidentyfikowali dodatkowe cechy, które również wykazują pozytywną korelację z **wymiarem aktywacji**, są to:

- zakres zmian wysokości tonu (pitch range),
- energia w wysokim paśmie częstotliwości (high-frequency energy),
- późne maksima w w przebiegu intensywności (late intensity peaks),
- wzrost intensywności w przebiegu jednostek znaczących (3-4-wyrazowe frazy wyodrębione z wypowiedzi),
- stromość wzrostów w przebiegu parametru F0 między maksimami sylab (slope of F0 rises),
- krótsze pauzy oraz krótszy iloczas fragmentów wypowiedzi między pauzami lub grupami oddechowymi (co ma związek z przyspieszoną aktywnością oddechową)

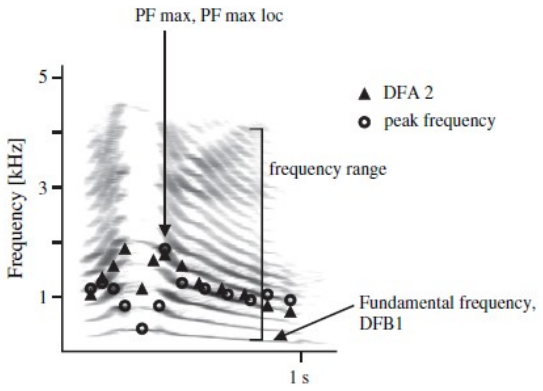
W odniesieniu do **wartościowości i wpływu** zidentyfikowano znacznie mniej cech akustycznych, które opisują te wymiary, być może z tej przyczyny, że brano pod uwagę zbyt małe zbiory parametrów lub dlatego, że aspekty ekspresji głosowej emocji związane z tymi wymiarami są przekazywane za pomocą tych samych cech akustycznych co aktywacja (odnosi się to szczególnie to wymiary wpływu/kontroli). Cechy, które w poprzednich badaniach wykazały związek z **wartościowością** to głównie cechy prozodyczne:

- tempo wypowiedzi,
- energia w wysokim paśmie częstotliwości (high-frequency energy),
- zakres zmian wysokości tonu,
- średnia wysokość tonu,
- dłuższy iloczas samogłoskowy oraz
- brak wzrostu w przebiegu intensywności w obrębie jednostek znaczeniowych.

wymiarach (->SEC)

Wyniki pracy Patel et al. (z 2011 roku, opisane w następnej sekcji) pokazują, że **wartościowość** jest prawdopodobnie związana raczej z mimiką twarzy i artykulacją niż z cechami prozodycznymi lub wynikającymi z **jakości głosu**.

W pracy **Hammerschmidt'a & Juergens'a** szukano akustycznych korelatów 6 emocji: wściekłość, rozpacz, wstręt (emocje „nieprzyjemne", negatywne w wymiarze *wartościowości*), radosne zaskoczenie (joyful surprise, coś pomiędzy euforią – elation a zdziwieniem – astonishment), sympatia/czułość i zmysłowe zadowolenie (emocje pozytywne w wymiarze wartościowości, określone przez autorów jako „przyjemne").
Materiał językowy stanowiły jednowyrazowe wypowiedzi (imię) zrealizowane dwukrotnie przez 23 mówców w każdym ze stanów emocjonalnych (23x2x6). W badaniu percepcyjnym oceniono, czy emocje zostały „właściwie zrealizowane". Wyniki pokazały, że na podstawie cech głosu mówców można było z dokładnością większą niż przypadkowe prawdopodobieństwo (16,7%) określić ich stan emocjonalny. Dokładność ta była jednak różna dla różnych stanów emocjonalnych (największa dla wściekłości, najmniejsza dla wstrętu).
W pierwszym podejściu dokonano ekstrakcji 18 parametrów dla segmentów tonalnych (only periodic peaks) i 76 dla segmentów tonalnych i szumowych. Po zbadaniu korelacji wyłoniono zbiór 15 parametrów opisujących prozodię wypowiedzi nacechowanych różnymi emocjami.
Przeprowadzono analizę dyskryminacyjną, która pozwoliła zidentyfikować 6 najważniejszych cech (spośród 15, zaznaczone kolorem w tabeli) z punktu widzenia klasyfikacji emocji – poprawną klasyfikację osiągnięto w przypadku 77% wypowiedzi. Liczba parametrów, które miały istotny wkład w rozróżnienie pomiędzy parami emocji wahała się między 3 (np. sympatia/czułość– wstręt) a 14 (np. wściekłość – zmysłowe zadowolenie).
tabela + wyjaśnienie (frequency range – spectrum not F0)



Parameter	***Description***
Amplitude	Relative amplitude based on the logarithmic root mean square method
Duration (ms)	Time from onset to end of utterance
Noise (%)	Percent of time segments in which no clear harmonic structure could be detected

F0 mean (Hz)*	Fundamental frequency, mean across all time segments
F0 max (Hz)*	Maximum of fundamental frequency
PF mean (Hz)*	Frequency with the highest amplitude, mean across all time segments
PF max (Hz)*	Maximum peak frequency
PF/F0 coeff mean*	Relation between mean peak frequency and mean fundamental frequency [(PF2F0)/F0]
DFB ratio*	Amplitude ratio between lowest (DFB1) and second-lowest dominant frequency band (DFB2), mean across all time segments; DFB1 corresponds to F0 in tonal segments
HNR mean*	Difference between the amplitude peaks of the DFBs and the noise level at the same frequency point, mean across all time segments
Range mean (Hz)*	Difference between minimum and maximum frequency in the power spectrum of a time segment of 5 milliseconds
DFA2 mean (Hz)	Frequency at which the amplitude distribution reaches the second quartile, mean across all time segments
DFB1 local mod (%)	Frequency with which the original curve of DFB1 crosses the average level of DFB1 (percentage of all time segments)
PF amp max (Hz)	Frequency of the maximum amplitude of the peak frequency
PF max loc	Relative position of PF max in the utterance (1/duration [ms] 3 time of PF max [ms] from voice onset;

Błędne klasyfikacje wystąpiły najczęściej między parami emocji:
zmysłowe zadowolenie – wstręt, sympatia/czułość
radosne zaskoczenie – wściekłość, rozpacz

Ponieważ jednym z celów badania było wyłonienie akustycznych korelatów emocji „nieprzyjemnych" zbadano wpływ parametrów akustycznych na rozróżnienie pomiędzy wypowiedziami nacechowanymi emocjami takimi jak wściekłość, rozpacz i wstręt z jednej strony i zmysłowe zadowolenie, radosne zaskoczenie i sympatia/czułość z drugiej (emocje „przyjemne").
Okazało się, że nośnikiem informacji o wartościowości emocji (tutaj: przyjemne vs. nieprzyjemne) są następujące parametry akustyczne: *amplituda, iloczas, PFmean, PFmax, PF/F0 coeff mean, DFA2 mean*. Wartości tych parametrów są wyższe w przypadku emocji „nieprzyjemnych" (wyjątek stanowi iloczas, który wykazuje odwrotną zależność), a ponadto wypowiedzi o tego rodzaju nacechowaniu emocjonalnym charakteryzują się większym *zaszumieniem* (noise%), *szerszym zakresem częstotliwości* (range mean) oraz *późniejszym*

maksimum częstotliwości (PFmax loc). Udział poszczególnych parametrów w rozróżnieniu między klasami emocji jest różny: najlepszym korelatem wartościowości emocji jest *PF/F0 coeff mean* a najsłabszym *DFA2 mean*.

Wyniki te nie potwierdzają się w przypadku analizy emocji o przeciwnej wartościowości w parach, co jest wynikiem wpływu głośności (po wyelimiowaniu tego wpływu uzyskano wyniki takie jak w analizie w grupach).

Kierunek zmian w wartościach parametrów akustycznych opisujących prozodię wypowiedzi nacechowanymi emocjami „głośnymi" („loud prosodies" – wściekłość & radosne zaskoczenie) a o średnim (rozpacz) lub niskim poziomie amplitudy („low-voice prosodies", pozostałe) był podobne do tego, który można było zaobserwować między emocjami „nieprzyjemnymi" i „przyjemnymi". Jednak w przypadku rozróżnienia między „loud" a „low-voice" prosodies istotne zmiany dotyczyły wszystkich parametrów oprócz PFmax loc, natomiast związek między wzrostem intensywności a wartościami parametrów *duration, noise* i *HNR mean* okazał się nieliniowy.

Podsumowanie

W pracy zidentyfikowano 6 parametrów akustycznych pozwalających na poprawną klasyfikację 77% wypowiedzi emocjonalnych.

Najmniej błędnych klasyfikacji miało miejsce w przypadku wypowiedzi nacechowanych wściekłością (hot anger).

Najwięcej wspólnych cech akustycznych mają *wściekłość* i *radosne zaskoczenie*: względnie wysokie F0 i frequency range, niestabilne F0 (DFB1 local mod), wysokie DFB ratio i PFamp max. Co odróżnia te emocje to niższa amplituda, dłuższy iloczas i mniej energii w wyższych częstotliwościach spektrum (DFA2 mean) w przypadku *radosnego zaskoczenia.*

Największe różnice akustyczne zaobserwowano między wściekłością a zmysłowym zadowoleniem i sympatią/czułością.

Wyniki dotyczące akustycznych korelatów emocji nieprzyjemnych nie potwierdzają obserwacji poczynionych we wcześniejszych pracach, w których badano cechy sygnalizujące poziom stresu mówcy. W pracach tych zidentyfikowano liniowy związek między stresem a następującymi parametrami: średnie i maksymalne F0, F0 STD, amplituda. Wyniki przedstawione w omawianej pracy pokazały, że parametry te nie są jednak nośnikiem informacji o wartościowości emocji, gdyż ich wartości są wysokie w przypadku niektórych emocji „przyjemnych" jak i „nieprzyjemnych". Autorzy tłumaczą to faktem, że wszystkie parametry F0 są wysoko skorelowane z amplitudą, więc wzrost amplitudy (głośności) sygnału będzie powodował wrażenie wyższego tonu. Gdyby autorzy odnieśli się w swojej pracy do jakiegoś modelu emocji mogliby dojść do konkluzji, że brak związku między wysokością F0 i amplitudy a wartościowością emocji jest wynikiem błędnego pogrupowania emocji (przyjemne – nieprzyjemne), gdyż nie odnosi się ono w rzeczywistości do

żadnego z uznanych w nauce o emocjach wymiaru np. autorzy wcale nie odnoszą się do wymiaru pobudzenia ani wpływu, co mogłoby pomóc w wyjaśnieniu zaobserwowanych związków między parametrami a głosową ekspresją emocji.
Jednak sama aplituda także nie jest dobrym predyktorem wartościowości emocji. Ostatecznie ustalono (tj. biorąc pod uwagę emocje o tym samym stopniu głośności), że za akustyczny korelat emocji nieprzyjemnych można uznać: PF/F0 coeff mean, PF max, noise, PF mean, range mean, PFmax loc i DFA2 mean.

Rola jakości głosu (voice quality) w przekazywaniu informacji o stanie emocjonalnym mówcy

Większość prac poświęconych akustycznym korelatom emocji i leżących u ich podstaw wymiarów (tj. wartościowość, aktywacja, wpływ) nie miała solidengo oparcia w żadnej teorii emocji i dlatego wybór parametrów, które badano, był przypadkowy. Najwięcej uwagi poświęcono parametrom związanym z *częstotliwością podstawową, intensywnością* i *iloczasem*, między innymi dlatego, że można je bez problemu uzyskać za pomocą standardowych narzędzi analizy sygnału mowy. Mimo, że niektóre z nich okazały się przydatne w dyskryminacji ekspresji głosowej różnych emocji, to nie wyjaśniono leżących u jej podstaw mechanizmów. Ponadto, podstawowe parametry pozwalają na dyskryminację w obrębie małych zbiorów emocji i nie są efektywne na większych zbiorach (wtedy trzeba włączyć dodatkowe cechy). Fakt, że słuchacze są w stanie z prawie 100% dokładnością rozpoznać stan emocjonalny mówcy na podstawie samego tylko głosu świadczy o tym, że istnieją wzorce akustyczne charakterystyczne dla emocji, ale jak do tej pory nie udało się zidentyfikować cech, które się na nie składają. Skłoniło to badaczy do roszerzenia zestawu badanych cech akustycznych i uwzględnienia parametrów bardziej złożonych, głównie związanych z opisem ilościowym jakości głosu (rozumianej jako fizyczne zmiany w drganiu więzadeł głosowych i kształcie toru głosowego z wykluczeniem zmian wysokości tonu, głośności i kategorii fonetycznej). Jednym z ważniejszych akustycznych korelatów emocji okazał się być parametr opisujący aktywność więzadeł głosowych: stosunek energii w wyższej i niższej części LTAS[4]. Większość badań nie zmierzała jednak do ustalenia związku pomiędzy zmianami fizjologicznymi leżącymi u podstaw emocji a mechanizmami produkcji głosu. Próbę taką podjęto w pracy Patel et al, w której oparto

4rodzaj widma długookresowego; daje informację jaka część całkowitej energii sygnału jest przenoszona w konkretnym paśmie częstotliwości i określa średnią moc sygnału w funkcji częstotliwości

się na modelu procesów składowych (CPM) Scherer'a i uwzględniono wpływ czynników *push* i *pull* na ekspresję głosową emocji.

Chociaż jakość głosu ma podłoże fizjologiczne, a więc jest przejawem wpływu czynników *push*, to nie pozostaje bez wpływu czynników *pull*, który można wyeliminować poprzez odpowiedni dobór materiału. Z tego powodu w cytowanej pracy zamiast pełnych wypowiedzi (które zrealizowane zostałyby z jakąś prozodią – pull effects) posłużono się wybuchami afektu (affect bursts). 10 aktorów obu płci wyprodukowało głoskę „a" imitując każdy z pięciu stanów emocjonalnych: smutek, ulgę, radość, paniczny strach i gwałtowną złość. Zmierzono 12 parametrów: *jitter, shimmer, HNR, średnie F0, equivalent sound level*, oraz uzyskane na podstawie elektroglottogramów (za pomocą *inverse filtering*, które pozwala na wydobycie dźwięku wytwarzanego w głośni i nie zmodyfikowanego przez kanał głosowy): *pulse amplitude, maximum flow declination rate (MFDR), normalized amplitude quotient (NAQ), closed quotient, różnica poziomów między pierwszą i drugą harmoniczną (H1-H2), alpha ratio, H1-H2$_{LTAS}$.* Pomiary znormalizowano względem mówcy.

ANOVA wykazała statystycznie istotny wpływ ($p<0.05$) emocji na wartości wszystkich badanych parametrów oprócz NAQ.

Przeprowadzono test Bonferroniego (uwzględniono tylko parametry dla których znaleziono statystycznie istotne różnice), w którym analizowano wartości parametrów pochodzących z pomiarów **par** wypowiedzi nacechowanych emocjonalnie, np. zbadano czy wartości *jitter* są statystycznie istotnie różne w wypowiedziach wyrażających ulgę i radość, itp.

Wyniki pokazały, że niektóre z badanych cech akustycznych są przydatne w rozróżnianiu pomiędzy dwiema emocjami (np. closed quotient tylko ulga-strach) podczas gdy inne opisują różnice w ekspresji głosowej więcej niż jednej pary emocji (np. średnie F0: ulga-strach/radość/złość, smutek-strach/radość/złość, strach-złość).

Test pokazał również, że niektóre emocje różnią się znacznie pod względem cech akustycznych (a więc produkcji głosowej), np. strach i ulga (statystycznie istotne różnice znaleziono w wartościach 9 z 11 badanych parametrów.)

W następnej kolejności dokonano analizy głównych składowych (principal component analysis) aby parametry, które wykazują podobną zmienność opisać za pomocą pojedynczego składnika lub wymiaru. Uzyskano trzy składniki, które w 83,5% odpowiadają za różnice między emocjami:

Pierwszy składnik zawiera cechy (closed quotient, MFDR, H1-H2, H1-H2$_{LTAS}$) związane z **wysiłkiem fonacyjnym** (phonatory effort). Wysiłek fonacyjny jest wynikiem ciśnienia podgłośniowego, które zmienia się wraz z szybkością i głębokością oddechu, te zaś związane są bezpośrednio z pobudzeniem ergotroficznym. Wysiłek fonacyjny reprezentuje wymiar aktywacji (pobudzenia) w emocji (zob. punkt b) sekcja 1.5.2) i pozwala na odróżnienie ulgi od radości, strachu i złości.

Drugi składnik zawiera cechy (jitter, shimmer i HNR), które określają **zakłócenia** w pracy więzadeł głosowych (perturbation) będące wynikiem zmian w ich napięciu (albo zbyt duże, albo zbyt małe). Można uznać, że składnik ten odzwierciedla podwymiar *power subcheck* (sprawdzenie możliwości uwolnienia się spod dominacji bodźca poprzez działanie zewnętrzne) w wymiarze *coping potential* (zob. punkt d) sekcja 1.5.1). Dzięki pomiarowi zakłóceń drgania więzadeł głosowych można odróżnić złość i ulgę od pozostałych emocji.

Na trzeci komponent składa się wyłącznie średnia wartość F0, tak więc odzwierciedla on **częstotliwość fonacji** (phonation frequency). Zmiany F0 są związane z oceną kontroli bodźca (control subcheck) w wymiarze *coping potential:* w rozwoju filogenetycznym wysokie F0 charakteryzowało ekspresję głosową osobników słabszych, przekonanych o braku wpływu na wynik i konsekwencje zdarzenia oraz o niemożliwości uwolnienia się spod dominacji bodźca. Określenie częstotliwości fonacji pozwala na rozróżnienie między ulgą (niskie F0) i panicznym strachem (wysokie F0) a radością i złością (średnie F0).

Nie znaleziono akustycznych korelatów wymiaru wartościowości – być może jest to spowodowane tym, że ocena bodźca w tym wymiarze nie ma bezpośredniego wpływu na produkcję głosu, lecz wiąże się ze zmianami w kształcie toru głosowego (-> pomiar formantów). Być może też ważniejszym niż głos nośnikiem informacji o wymiarze wartościowości jest mimika twarzy.

Podsumowując, parametry opisujące cechy źródła głosu niosą istotne informacje na temat stanu emocjonalnego mówcy, a zmienność w ich obrębie jest uwarunkowana fizjologicznymi zmianami wywołanymi przez ocenę bodźca w kilku wymiarach. Wzorce akustyczne emocji składają się z trzech komponentów: wysiłku fonacyjnego, zakłócenia i częstotliwości fonacji, wyłonionych na podstawie analizy podobieństwa w zmienności przebiegu parametrów źródła głosu.

C1: wysiłek fonacyjny -> wymiar aktywacji

C2 & C3: zakłócenia & częstotliwość fonacji -> wymiar wpływu (kolejno *power* i *control subcheck*)

Rozpoznawanie mowy emocjonalnej dla systemów dialogowych

Uwzględnienie emocji w głosie w systemach dialogowych wpływa na ich naturalność i wydajność, np. systemy wykorzystywane w informacji telefonicznej będą mogły wygenerować odpowiedź dla użytkownika odpowiednią do jego stanu emocjonalnego lub przekazać

połączenie do konsultanta. Trzeba zauważyć, że w systemach dialogowych wcale nie jest konieczne uwzględnienie dużej liczby emocji. Wręcz przeciwnie, należy wziąc pod uwagę tylko te, które mają znaczenie z punktu widzenia zastosowania systemu. W odniesieniu do informacji telefonicznej opartej na systemie dialogowym naistotniejsze wydaje się rozróżnienie pomiędzy emocjami negatywnymi i nie-negatywnymi (negative and non-negative) ponieważ wykrycie negatywnych emocji pozwoli na przekazanie połączenia do konsultanta, co przyczyni się do polepszenia jakości świadczonych usług.

Wykorzystano nagrania pochodzące z bazy rozmów telefonicznych zarejestrowanych w informacji telefonicznej wykorzystującej system dialogowy. Dokonano analizy i klasyfikacji nagranych wypowiedzi pod kątem zawartości emocjonalnej. Do wypowiedzi o *negatywnym* zabarwieniu emocjonalnym zaliczono te, w których mówca przejawiał złość lub frustrację, natomiast do *nie-negatywnych* – wypowiedzi *neutralne* oraz o zabarwieniu emocjami *pozytywnymi* jak np. radość.

Wzięto pod uwagę **21 parametrów akustycznych** związanych z F0, energią, iloczasem i formantami. Pomiarów dokonano na długości całych wypowiedzi. Dla F0 i energii zmierzono: średnią, medianę, odchylenie standardowe, maksimum, minimum, współczynnik regresji liniowej oraz zakres (max-min). Parametry związane z iloczasem to: tempo wypowiedzi, stosunek iloczasu fragmentów dźwięcznych do bezdźwięcznych, iloczas najdłuższego fragmentu dźwięcznego w wypowiedzi. Określono średnie wartości częstotliwości F1 i F2 i szerokości ich wstęg.

Wybrano 2 zestawy cech (10 i 15 parametrów, forward feature selection), które badano osobno dla mówców żeńskich i męskich. Wśród najważniejszych cech dla obu płci znalazły się:

- stosunek iloczasu fragmentów dźwięcznych do bezdźwięcznych
- mediana energii
- współczynnik regresji liniowej F0

Wyłoniono także zbiór najlepszych korelatów akustycznych emocji za pomocą PCA.

Oprócz cech akustycznych wykorzystano parametry niosące **informacje leksykalne** oraz **dyskursywne**. Informacja **leksykalna** odnosi się do *emocjonalnej istotności* wyrazów (emotional salience). Wyrazy w wypowiedzi mogą mieć różne znaczenie dla przekazu stanu emocjonalnego mówcy: te o większym znaczeniu mają wyższy wskaźnik istotności emocjonalnej.

Na poziomie **dyskursu**, wypowiedzi użytkownika przypisano do jednej z **kategorii**:

- odmowa (rejection),
- powtórzenie,
- przeformułowanie,
- ask-start over (gdy użytkownik prosił o pomoc lub chciał wrócić na początek rozmowy),
- żaden z powyższych (tutaj znalazły się odpowiedzi użytkownika proszonego o podanie konkretnych informacji np. osobowych).

Do klasyfikacji wykorzystano dwa rodzaje modeli: LDC (linear discrimination classifiers) oraz k-NN (k-nearest neighbourhood classifiers). Najlepsze wyniki klasyfikacji za pomocą LDC uzyskano kolejno na następujących zbiorach cech:

głos męski	głos żeński
ac+lan, PCA	ac+lan, f10
ac+lan, f10	ac+lan, f15
ac+lan, f15	ac+lan, PCA
ac+lan+dis, base	ac+lan, base

i dla modelu k-NN

głos męski	głos żeński
ac+lan, f10	ac+lan, f15
ac+lan, f15	ac+lan, base
ac+lan, base	ac+lan, f10
ac+lan, PCA	ac+lan, PCA

W porównaniu z wynikami otrzymanymi dla modeli wykorzystujących wyłącznie informację akustyczną po dodaniu informacji leksykalnej (istotność emocjonalna wyrazów) uzyskano w niektórych przypadkach 80% redukcję błędu klasyfikacji. Dodanie informacji dyskursywnej nie wpłynęło na wyniki ze względu na wysoką korelację z informacją leksykalną.
Na podstawie tych wyników można wywnioskować, że podczas gdy cechy akustyczne (głównie związane z prozodią wypowiedzi) niosą informacje o wymiarze aktywacji, to wymiar wartościowości jest być może kodowany na poziomie leksykalnym, stąd znacząca rola emocjonalnej istotności wyrazów w rozróżnianiu między wypowiedziami nacechowanymi emocjami negatywnymi a tymi, które takiego zabarwienia nie mają.

6.2. Badania zorientowane na percepcję

Baenziger & Scherer (2003) podjęli próbę zidentyfikowania percepcyjnego cech głosu, które niosą informacje na temat stanu emocjonalnego mówcy i powiązania ich z parametrami akustycznymi.
Słuchaczom zaprezentowano „bezsensowne” wypowiedzi o różnym zabarwieniu emocjonalnym: zimna złość (irytacja) – gwałtowna złość (wściekłość), niepokój – paniczny strach, smutek – rozpacz, szczęście – euforia. (Pary tych emocji różnią się między sobą stopniem aktywacji).

W specjalnym programie słuchacze mieli dokonać oceny percepcyjnej następujących aspektów wypowiedzi emocjonalnych:

- wysokość tonu (niski – wysoki)
- intensywność (mała – duża)
- intonacja (monotonna – zróżnicowana)
- tempo wypowiedzi (szybkie – wolne)
- artykulacja (dokładna – niedokładna)
- stabilność głosu (głos stabilny lub trzęsący się; instability – shaky or steady)
- szorstkość głosu (roughness)
- ostrość głosu (sharpness)

Zaobserwowano, że w stosunku do niektórych cech, np. intensywności odpowiedzi słuchaczy były bardziej zgodne niż w w stosunku do innych, np. artykulacji lub szorstkości głosu. Wykonano analizę czynnikową, w której wyłoniono **2 komponenty**, które odpowiadają w 73% za percepcyjne rozróżnienie między emocjami. Pierwszy z nich zawiera **cechy prozodyczne**: intenswyność, intonację, ostrość, wysokość tonu i tempo wypowiedzi. Na drugi komponent składają się stabilność, artykulacja i szorstkość, można więc powiedzieć, że reprezentuje on wymiar **jakości głosu**. Dokonano pomiarów parametrów akustycznych na analizowanych percepcyjnie wypowiedziach. Wzięto pod uwagę parametry związane z F0, intensywnością, iloczasem i dystrybucją energii spektralnej. Zbadano korelację między nimi i wybrano 10 względnie niezależnych parametrów, które odgrywały znaczną rolę w jednym z 2 komponentów (prozodycznym lub związanym z jakością głosu):

- F0 min i zakres,
- zakres intensywności
- całkowity iloczas
- stosunek iloczasu segmentów dźwięcznych do całkowitego
- ilość energii segmentów dźwięcznych w zakresie a) 300-500Hz i b) 600-800Hz
- ilość energii poniżej 1kHz we fragmentach a) dźwięcznych i b) bezdźwięcznych
- średnia intensywność

Związek pomiędzy 10 parametrami akustycznymi a wymiarami zmian głosu badanymi percepcyjnie przedstawia się następująco (+ ozn. pozytywną korelację, a – negatywną):

vocal dimension	***acoustic parameters***	***R2***
intensity	int.mean(+), int.range(+), v.0-1k(-)	0.88
sharpness	int.mean(+), F0.range(+), F0.min(+),	

	v.0-1k(-),int.range(+)	0.87
speed	dur.tot(-), int.mean(+), v.0-1k(-), dur.v/art(-)	0.79
intonation	F0.range(+), int.mean(+), int.range(+), F0.min(+),	
	n.0-1k(-), dur.tot(-)	0.67
pitch	F0.min(+), F0.range(+), int.mean(+)	0.65
instability	F0.min(+), dur.tot(+), v.0-1k(+)	0.35
articulation	F0.min(-), int.mean(+), int.range(+), F0.range(-),	
	n.0-1k(-)	0.32
roughness	n.0-1k(+), v.0-1k(-), int.mean(+), dur.tot(+)	0.28

Na podstawie tych wyników możemy stwierdzić, że średnia intensywność jest najlepszym predyktorem percepcyjnej intensywności oraz ostrości głosu, ale ma również duże znaczenie w percepcji innych wymiarów: tempa (speed), intonacji, wysokości tonu, artykulacji oraz szorstkości głosu. F0 min jest najlepszym akustycznym korelatem percepcyjnych zmian wysokości tonu, stabilności głosu oraz artykulacji. Percepcyjne zmiany w intonacji są sygnalizowane akustycznie przede wszystkim przez zakres F0 (F0 range). Spośród parametrów opisujących dystrybucję energii istotne są ilość energii poniżej 1kHz we fragmentach dźwięcznych (v.0-1k; wszystkie wymiary percepcyjne oprócz intonacji, tonu i artykulacji) i bezdźwięcznych (n.0-1k; intonacja, artykulacja i szorstkość głosu – najważniejszy korelat). Kolejny wniosek dotyczy stopnia w jakim parametry akustyczne przyczyniają się do percepcyjnej oceny zmian w poszczególnych wymiarach: w przeciwieństwie do wymiarów związanych z intonacją, w odniesieniu do jakości głosu (wymiary: instability, articulation, roughness) udział wziętych pod uwagę parametrów jest w tym względzie niewielki.

Zaobserwowano również, że w przypadku wypowiedzi nacechowanych emocjami o niskim stopniu aktywacji (radość, smutek, niepokój, zimna złość/irytacja) tempo jest pozytywnie skorelowane z dokładnością artykulacji: wypowiedzi, które zostały zrealizowane w szybszym tempie (kolejno: irytacja, niepokój, radość) cechują się także dokładniejszą artykulacją (w ocenie percepcyjnej słuchaczy) niż te wolniejsze (smutek). Ogólnie rzecz biorąc można wysunąć hipotezę, że wypowiedzi nececchowane smutkiem i irytacją są realizowane w innym stylu (slurred vs. controlled speaking style).

Stworzono również model, który pozwolił na ocenę wkładu cech akustycznych z jednej strony i percepcyjnych wymiarów z drugiej do komunikacji głosowej emocji. Wyniki pokazały, że we wszystkich stanach emocjonalnych charakteryzujących się niskim stopniem aktywacji, a szczególnie w przypadku *radości,* percepcyjne wymiary w znacznie większym stopniu wyjaśniały związek między wyrażaną (expressed) a postrzeganą (perceived) emocją niż wyłonione w analizach statystycznych cechy akustyczne. Spośród badanych za pomocą modelu

emocji, cechy akustyczne najlepiej opisują związek między wyrażaną i postrzeganą irytacją.

Wnioski

a) konieczność uwzględnienia bardziej złożonych parametrów akustycznych

b) znalezienie takich parametrów, które będą wysoko skorelowane z wymiarami percepcyjnymi, które w istotny sposób przyczyniają się do percepcji i ekspresji emocji

1.2. Badania zorientowane na produkcję i percepcję

W obszernym opracowaniu Banse & Scherer'a pt. „Acoustic profiles of vocal emotion expression" (1996 r.) szukano odpowiedzi na dwa pytania:

1. Czy słuchacze na podstawie samego sygnału mowy są w stanie rozpoznać stan emocjonalny mówcy?
2. Jakie są specyficzne wzorce ekspresji głosowej dla konkretnych emocji?

Jak zauważają autorzy, wyniki wcześniejszych prac (tzw. decoding studies) pokazały, że słuchacze są w stanie z ok. 50% dokładnością określić stan emocjonalny mówcy na podstawie jego głosu, co przewyższa 4-5-krotnie dokładność przypadkowego rozpoznawania (tj. gdy wskazana zostaje najczęstsza kategoria). Jednakże nie wszystkie emocje są rozpoznawane z jednakową dokładnością, np. smutek lub złość są znacznie częściej poprawnie klasyfikowane na podstawie głosu niż wstręt.

1. Wybór emocji

W pracy Banse & Scherer'a uwzględniono tylko te emocje, które w poprzednich badaniach słuchacze identyfikowali z dużą dokładkością i posłużono się licznym zbiorem emocji (14), aby wykluczyć możliwość, że słuchacze zamiast rozpoznawać emocje będą dokonywali ich klasyfikacji poprzez dyskryminację, co może mieć miejsce gdy zbiór jest nieliczny (<10). Kolejnym kryterium wyboru emocji było uwzględnienie różnic między nimi związanych z wymiarem aktywacji i jakości (quality),

dlatego też posłużono się 7 parami emocji reprezentujących różne rodziny emocji (emotion families, Eckman 1992):

zimna złość (irytacja) i gwałtowana złość (cold vs. hot anger)

radość i euforia,

smutek i rozpacz

niepokój i paniczny strach

duma i wstyd

zainteresowanie i nuda

pogarda i wstręt

2. Materiał

Materiał badawczy stanowiły nagrania 2 „nieznaczących" wypowiedzi (dwie różne sekwencje sylab z różnych języków – kotrola czynników *pull*), zrealizowanych przez 12 aktorów (6 k, 6 m) dwukrotnie w każdym ze stanów emocjonalnych. Dodatkowo, biorąc pod uwagę fakt, że dany stan emocjonalny może powstać pod wpływem różnych czynników/zdarzeń (tzw. antecedent situations/events), dla każdej z 14 emocji przygotowano 2 scenariusze prezentujące typowy dla niej kontekst. W sumie nagrano 1344 (2x12x14x2x2) wypowiedzi.

Wstępna ocena nagrań pozwoliła stwierdzić, że nie wszyscy aktorzy jednakowo „dobrze" zrealizowali zadanie. Tak więc, aby mieć pewność, że nagrania faktycznie są reprezentatywne dla wybranych stanów emocjonalnych dokonano ich selekcji, która polegała na ocenie **autentyczności i rozpoznawalności emocji** na podstawie: samego sygnału, samego nagrania video, audio+video. Oceny dokonało 12 studentów szkoły aktorskiej.
Spośród wszystkich 1344 wypowiedzi wyłoniono 244 pod kątem wymienionych kryteriów oraz w taki sposób, że materiał końcowy zawierał wszystkie 14 emocji zrealizowanych dwukrotnie w każdym ze zdań w dwóch scenariuszach, raz przez mówcę męskiego i drugi – przez mówcę męskiego (14x2x2x2x2).
Wyniki pokazały, że dystrybucja wypowiedzi między aktorami jest bardzo nierówna, gdyż 88% wyselekcjonowanych wypowiedzi pochodziło od 3 mówców.

3. Test percepcyjny

Te 224 wypowiedzi zostały następnie poddane percepcyjnej ocenie przez „naiwnych" słuchaczy, którzy mieli określić stan emocjonalny mówcy na podstawie samego tylko głosu.
Wyniki testu pokazały, że na podstawie samego tylko głosu słuchacze są w stanie określić stan emocjonalny mówcy z dużą dokładnością, tj. **48%** (55% gdy emocje reprezentujące tę samą rodzinę były traktowane jako pojedyncza kategoria, np. hot i cold anger; prawdopodobieństwo prawidłowej odpowiedzi przy przypadkowym wyborze wynosiło 100/14=7%). Wynik ten jest zbliżony do wyników prezentowanych we wcześniejszych pracach tj. ok. 50% dokładność rozpoznawania.
Trzeba zauważyć, że wynik Banse & Scherer'a może być zawyżony z uwagi na to, że materiał wykorzystany w eksperymencie został poddany wstępnej selekcji (co ograniczyło ilość mniej autentycznych realizacji wypowiedzi nacechowanych emocjonalnie), ale z drugiej strony równoważy to fakt, że autorzy wykorzystali znacznie większą liczbę emocji (14) niż przeciętnie, a wiadomo, że im mniej klas do rozpoznania tym łatwiejsze zadania (-> dyskryminacja zamiast rozpoznawania) i lepsze wyniki.

Nie wszystkie emocje są rozpoznawane z jednakową dokładnością. Niekoniecznie ma związek z realizacją wypowiedzi nacechowanych tymi emocjami przez aktorów, a raczej świadczy o tym, że:
1. wzorce akustyczne charakteryzujące ekspresję głosową tych emocji są mniej wyraziste ponieważ odzwierciedlają różnorodność związanych z nią modalności (jest bardzo wiele i bardzo różnych kontekstów – antecedent situations/events -> appraisals, w których osoba może zacząć odczuwać wstręt, np. coś brzydko pachnie, wygląda albo jest niezgodne z naszą postawą moralną);
2. emocje te nie są wyrażane w jednakowy sposób tj. nie we wszystkich stanach emocjonalnych mówca jest skłonny produkować wypowiedzi w formie **zdań**, w niektórych być może zamiast budować zdania będzie komunikował się za pomocą **wybuchów** lub **wzorców afektu** (affect bursts or emblems).
Najbardziej „rozpoznawalne" okazały się: **gwałtowna złość** (hot anger), **zainteresowanie, znudzenie** i **pogarda** (wszystkie powyżej **60%**).
Najtrudniejsze do sklasyfikowania okazały się **wstyd i wstręt**, których dokładność rozpoznania wyniosła odpowiednio **22% i 15%,** tj. tylko niewiele więcej niż przypadkowa klasyfikacja (7%). Równie słabe wyniki w rozpoznawaniu wstrętu osiągnięto w innych badaniach (np. 28% przy kilku zaledwie emocjach), co sugeruje, że prawdopodobnie naturalnie występujące ekspresje głosowe wstrętu składają się z wybuchów afektu (np. „ble" , ang. „yuck") a nie ze zdań realizowanych z charakterystycznym dla wstrętu wzorcem akustycznym.
W odniesieniu do wstydu, być może nie istnieją wzorce akustyczne charakterystyczne dla ekspresji tej emocji, ponieważ osoba, która się wstydzi chętniej będzie milczała, niż cokolwiek mówiła. Zarówno wstręt jak i wstyd są prawdopodobnie kodowane w znacznie większym stopniu przez zmiany w kanale wzrokowym (visual cues) niż głosowym (vocal cues).

Dla części emocji, uwzględnienie emocji należących do jednej rodziny jako jednej kategorii (np. smutek i rozpacz) wpłynęło pozytywnie na statystyki rozpoznania – tak było w przypadku hot & cold anger, panic fear & anxiety, desperation & sadness, przy czym odsetek poprawnie rozpoznanych emocji wzrastał w odniesieniu do obu emocji (tj. i tej mniej, i tej bardziej intensywnej). W odniesieniu do pozostałych, taki „zabieg" nie przyniósł żadnych efektów.

4. Parametry akustyczne
    a. Uwagi wstępne – dane z literatury

Choć do tej pory nie udało się jednoznacznie zidentyfikować akustycznych wzorców charakteryzujących ekspresję głosową różnych emocji to wiadomo, że udział w niej biorą następujące zmienne akustyczne:

- poziom, zakres i (w mniejszym stopniu) kontur F0

- energia (lub amplituda, postrzegana jako intensywność głosu)
- aspekty związane z przebiegiem czasowym wypowiedzi: tempo, pauzy itp.
- dystrybucja energii w spektrum, szczególnie stosunek względnej energii w wysokim paśmie częstotliwości względem niskiego (wpływa na precepcję jakości głosu, jego barwy)
- pozycje formantów (odzwierciedlają zmiany w kształcie kanału głosowego -> percepcja artykulacji)

Związek między tymi parametrami a emocjami przedstawia się następująco:

Złość: wzrost średniego F0 i energii. W badaniach, w których prawdopodobnie wzięto pod uwagę „hot anger" zaobserwowano także większy zakres i zmienność F0; w badaniach, w których nie stwierdzono takiej zależności badano prawdopodobnie „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 (także w „łagodniejszych" formach tej emocji jak , zmartwienie, niepokój) i jego zakresu, wysokiego poziomu energii w wyższym paśmie częstotliwości (high-frequency energy) oraz przyspieszonej artykulacji.
Smutek: jest emocją o niskim stopniu aktywacji, więc można spodziewać się spadków zamiast wzrostów. Dotyczy to średniego F0 i jego zkaresu, średniej energii, niższego poziomu energii w wyższym paśmie częstotliwości i wolniejszej artykulacji oraz konturów opadających (downward-directed F0 contours).
Jednak bardziej 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 energii, a także podwyższonej energii w wyższym paśmie częstotliwości i szybszej artykulacji.
Wstręt: wyniki analiz korelatów akustycznych wstrętu są niejednoznaczne (patrz uwagi powyżej). Różnice w wynikach można wyjaśnić przede wszystkim rodzajem materiału, którym posługiwano się w analizach. W analizach opartych na wzbudzaniu wstrętu poprzez pokaz specyficznych filmów zaobserwowano wzrost średniego F0, zaś w „portretach emocji" zrealizowanych przez aktorów – spadek średniego F0.

Intonacja – emocje. Czy istnieją wzorce intonayjne charakterystyczne dla konkretnych emocji?

Scherer zbadał to zagadnienie wykorzystując niewielki korpus składający się z 2 „nieznaczących" wypowiedzi zrealizowanych przez 9 aktorów w 8 stanach emocjonalnych (łącznie 144 wypowiedzi, kolorem zaznaczono emocje o wyższym stopniu pobudzenia):
zimna złość (irytacja) i gwałtowna złość (hot/cold anger)
niepokój i paniczny strach
smutek i rozpacz
radość i euforia
Kontury intonacyjne wypowiedzi były kodowane w taki sposób, aby można było dokonać ich analizy ilościowej: dokonano stylizacji liniowej między punktami wyznaczającymi akcenty oraz końcowy spadek (final fall). F0 znormalizowano względem mówcy (Patterson & Ladd).
Najczęstszym wzorcem intonacyjnym zrealizowanym przez mówców były dwa roznąco-opadające akcenty (po jednym we frazie), po których następował spadek F0 na ostatniej sylabie w wypowiedzi.

Liczba wzrostów i spadków F0 oraz akcentów w pierwszej i drugiej frazie wypowiedzi oraz na końcowej sylabie wykazały związek z mówcą oraz wypowiedzią (-> dwie różne sekwencje sylab), ale nie z emocją.

Wyniki można podsumować następująco:

W badanym materiale emocje wpłynęły głównie na wysokość i zakres F0, i te dwa parametry wystarczają aby podsumować najważniejsze różnice między emocjami: ogólnie rzecz biorąc emocje charakteryzujące się wysokim stopniem pobudzenia mają wyższe średnie F0 i szerszy zakres niż odpowiadające im emocje o niskim stopniu pobudzenia
W wypowiedziach nacechowanych odmiennymi emocjami zobserwowano odmienną realizację akcentów (wysokość maksimum, wielkość wzrostu i spadku F0) oraz różnice w kształcie konturu intonacyjnego. Kontury rosnące (*uptrend*: stopniowy wzrost poziomu F0 aż do final fall), wystąpiły w wypowiedziach nacechowanych rozpaczą i euforią, zaś odwrotny przebieg konturu F0 (*downtrend*: wczesne maksimum po którym następuje stopniowy spadek poziomu F0 aż do final fall) charakteryzował smutek i radość.
Przebieg final fall również wykazał związek z emocją: w przypadku gwałtownej złości i euforii był on realizowany przez większy spadek F0 niż w przypadku niepokoju i radości.
Jednak z uwagi na bardzo ograniczony materiał badawczy oraz dużą zmienność wewnątrz

analizowanych cech (duże odchylenia standardowe) należy stwierdzić, że nie udało się zidentyfikować konturów intonacyjnych charakterystycznych dla wybranych emocji. Przemawia za tym również fakt, że posłużono się pozbawionymi znaczenia leksykalnego wypowiedziami, w których wpływ ograniczeń składniowych i semantycznych oraz czynników pull był znacznie ograniczony.
Można więc przypuszczać, że czynniki push wpływają jedynie na ogólny poziom i zakres F0. Poza tym, badania percepcyjne z wykorzystaniem wypowiedzi, które poddano różnego rodzaju manipulacji (**random-splicing**: brak intonacji ale zachowane voice quality vs. **content-filtering**: maskowanie voice quality ale intonacja zachowana) sugerują, że **związek między emocjami a intonacją ma drugorzędne znaczenie dla rozpoznawania emocji w stosunku do ogólnego poziomu i zmienności F0 oraz aspektów spektralnych jakości głosu.**

Mozzicionacci & Hermes przeprowadzili badania w celu podobnym jak Scherer tj. ustalenia związku między intonacją a ekspresją emocji i nastawienia. Wykonali analizy pod kątem produkcji głosowej oraz precepcji. (dokończyć)

Załączniki do części II

Załącznik nr 1

Podsumowanie cech akustycznych różnych typów ekspresji głosowej

szeroki głos *(wide voice)*	✓ obniżona częstotliwość i zwiększenie szerokości pasma F1 ✓ więcej energii w niższych pasmach częstotliwości, ✓ nosowość gardłowo-językowa
wąski głos *(narrow voice)*	✓ więcej energii w wyższych pasmach częstotliwości, ✓ zmniejszenie szerokości pasm formantów (szczególnie F1) ✓ zmiany w częstotliwościach formantów (F1>, F2 i F3<) ✓ nosowość krtaniowo-gardłowa
napięty głos *(tense voice)*	✓ zmniejszenie szerokości pasm formantów (szcz. F1). ✓ podwyższone F0 i amplituda ✓ nieregularne i aperiodyczne drgania -> jitter i shimmer ✓ zwiększenie energii w wysokim paśmie częstotliwości spektrum (500-1000Hz) ✓ wrażenie szorstkości głosu
rozluźniony głos *(relaxed voice)*	✓ poziom F0 bliski granicy fizjologicznej (raczej niski ton głosu) ✓ poziom amplitudy w zakresie od niskiego do umiarkowanego ✓ brak jitter i shimmer, ✓ brak wrażenia szorstkości głosu, ✓ równomiernie rozłożona energia spektralna (możliwy jedynie niewielki spadek energii w wyższych partiach spektrum), ✓ brak zmian w częstotliwościach i szerokościach pasm formantów
całkowicie rozluźniony głos *(lax voice)*	✓ niskie F0 i zakres F0, ✓ niska amplituda, ✓ szum widmowy, ✓ nosowość, ✓ bardzo niski poziom energii w wysokich pasmach częstotliwości, ✓ słaba pulsacja, ✓ częstotliwości formantów bliskie wartościom neutralnym (-> niedokładna artykulacja), ✓ szerokie pasmo F1 ✓ breathy voice
pełny głos *(full voice)*	✓ rejestr piersiowy (chest register phonation), ✓ niskie F0 ✓ wysoka amplituda ✓ wysoka energia sygnału w całym zakresie częstotliwości
cienki głos	✓ rejestr głowowy

(thin voice)	✓	wysokie F0
	✓	niska amplituda
	✓	spektrum z szeroko rozmieszczonymi harmonicznymi
	✓	wyższy poziom energii wyższych składowych harmonicznych.

Załącznik nr 2

Wyniki ocen w wymiarach modelu kompozycyjnego Scherera w poszczególnych stanach emocjonalnych (za *Scherer, Vocal affect expression*)

			Goal/need significance			
Emotional state	Novelty	Pleasantness	Relevance	Expectation	Conduciveness	Urgency
Enjoyment/ happiness	Low	High	Medium	Consistent	High	Very low
Elation/joy	High	High	High	Discrepant	High	Low
Displeasure/ disgust	Open	Very low	Low	Discrepant	Low	Medium
Contempt/scorn	Open	Low	Low	Discrepant	Low	Low
Sadness/ dejection	Low	Low	High	Discrepant	Obstruct	Low
Grief/ desperation	High	Low	High	Discrepant	Obstruct	High
Anxiety/worry	Low	Open	Medium	Discrepant	Obstruct	Medium
Fear/terror	High	Low	High	Discrepant	Obstruct	Very high
Irritation/cold anger	Low	Open	Medium	Discrepant	Obstruct	Medium
Rage/hot anger	High	Open	High	Discrepant	Obstruct	High
Boredom/ indifference	Very low	Open	Low	Consistent	Obstruct	Low
Shame/guilt	Low	Open	High	Discrepant	Obstruct	Medium

	Coping potential			Norm compatibility	
Emotional state	Control	Power	Adjustment	External	Internal
Enjoyment/ happiness	—	—	High	High	High
Elation/joy	—	—	Medium	High	High
Displeasure/ disgust	Open	Open	High	Low	—
Contempt/scorn	Open	High	High	Low	—
Sadness/ dejection	None	—	Medium	—	—
Grief/ desperation	Low	Low	Low	—	—
Anxiety/worry	Open	Low	Medium	—	—
Fear/terror	Open	Very low	Medium	—	—
Irritation/cold anger	High	Medium	High	Low	Low
Rage/hot anger	High	High	High	Low	Low
Boredom/ indifference	Medium	Medium	High	—	—
Shame/guilt	High	Open	Medium	Very low	Very low

III. Przegląd baz danych z uwzględnieniem możliwości zastosowania w systemach weryfikacji i identyfikacji mówcy

1. Bazy danych ukierunkowane na zastosowanie w tworzeniu/testach systemów weryfikacji i identyfikacji mówców

1.1. Baza *Polycost*

Polycost - baza średnich rozmiarów, przygotowana z przeznaczeniem do porównywania i walidacji algorytmów identyfikacji mówcy, zawiera nagrania mowy czytanej z telefonu, wypowiedzi w języku angielskim oraz w języku ojczystym mówcy. W bazie Polycost zawarte są również nagrania dla j. polskiego.

- **dostępność**: katalog ELRA, od 500 € do 1200 € zależnie od zastosowań i czlonkostwa w ELRA
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=436

Korpus nagrano od stycznia do marca 1996 roku w ramach projektu "Speaker Recognition in Telephony" (COST 250). Dane zebrano z międzynarodowych linii telefonicznych, zawierają one minimum 5 sesji na mówcę. Każdy mówca wypowiada się w języku angielskim (w większości przypadków - jako obcym) oraz w języku ojczystym (zob. niżej). Korpus oferowany w katalogu ELRA zawiera 1 285 rozmów telefonicznych (ok. 10 sesji na mówcę) nagranych przez 133 mówców (74 mężczyzn, 59 kobiet) z 13 krajów, 17 języków. Większość krajów reprezentowana jest przez 10 mówców. Każda sesja zawiera 15 wypowiedzi: 1 wypowiedź dla detekcji DTMF (Dual Tone Multi Frequency), 10 sekwencji cyfr w języku angielskim, 2 zdania w j. angielskim, 2 zdania w języku ojczystym mówcy. Jedna z wypowiedzi w języku ojczystym to swobodna wypowiedź ('free speech').

W publikacjach na temat korpusu Polycost pojawia się m.in.informacja, że występują tam dane dla j. polskiego, ale tylko dla jednego głosu (np. Hennebert et al., 2000).

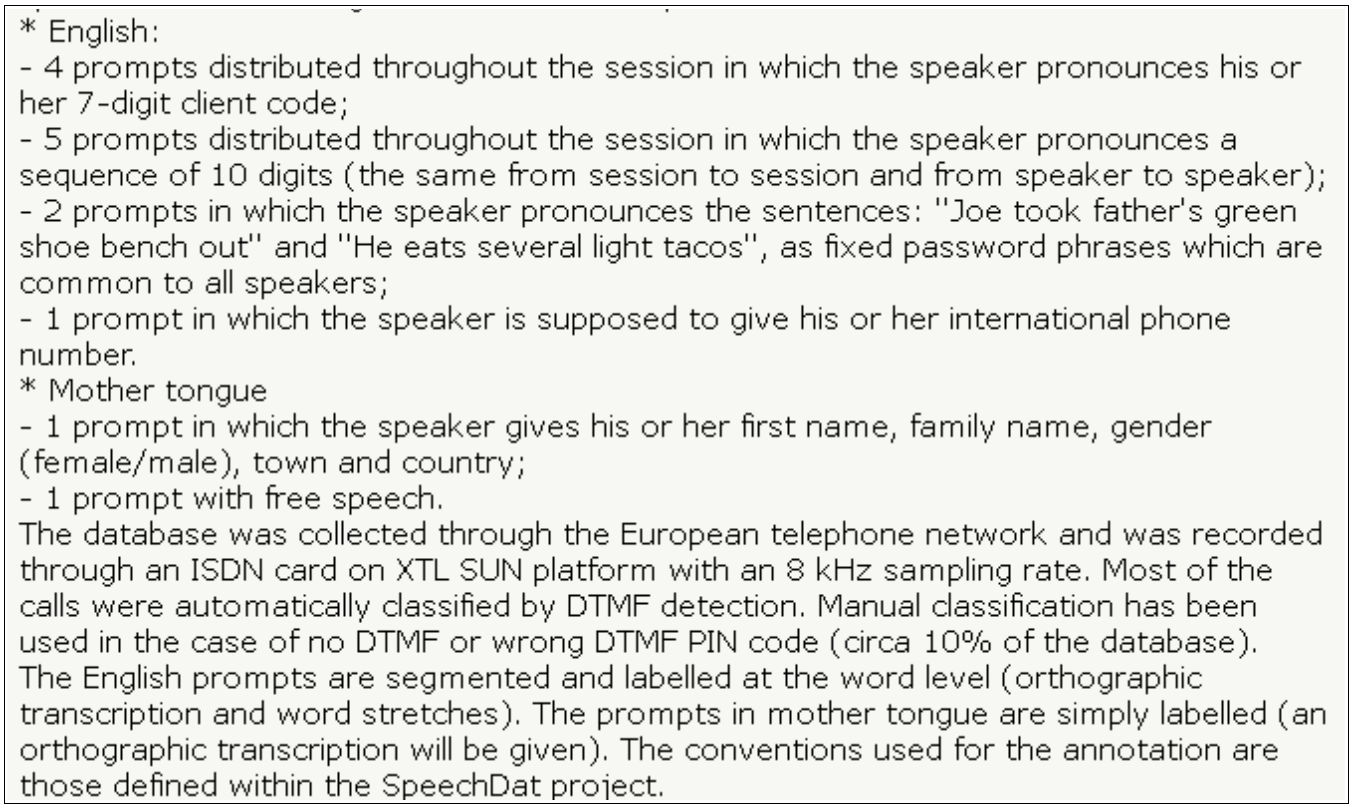
Ogólna liczba nagrań dla poszczególnych języków podsumowana została w tabeli wyżej.

Szczegółowe informacje także na stronie:

http://www.sciencedirect.com/science/article/pii/S0167639399000825#sec2

język	głosy męskie	głosy kobiece
włoski	5	5
portugalski	3	2
kataloński	2	1
hiszpański	3	4
francuski	23	16
szwedzki	6	4
duński	5	5
galicyjski	1	0
holenderski	7	5
niemiecki	1	0
angielski	10	10
turecki	5	5
litewski	1	0
rosyjski	1	0
arabski	0	2
macedoński	0	1
polski	1	0

Tabela 1. Polycost - języki i liczba mówców (z: Hennebert et al., 2000)


* English:
- 4 prompts distributed throughout the session in which the speaker pronounces his or her 7-digit client code;
- 5 prompts distributed throughout the session in which the speaker pronounces a sequence of 10 digits (the same from session to session and from speaker to speaker);
- 2 prompts in which the speaker pronounces the sentences: "Joe took father's green shoe bench out" and "He eats several light tacos", as fixed password phrases which are common to all speakers;
- 1 prompt in which the speaker is supposed to give his or her international phone number.
* Mother tongue
- 1 prompt in which the speaker gives his or her first name, family name, gender (female/male), town and country;
- 1 prompt with free speech.
The database was collected through the European telephone network and was recorded through an ISDN card on XTL SUN platform with an 8 kHz sampling rate. Most of the calls were automatically classified by DTMF detection. Manual classification has been used in the case of no DTMF or wrong DTMF PIN code (circa 10% of the database).
The English prompts are segmented and labelled at the word level (orthographic transcription and word stretches). The prompts in mother tongue are simply labelled (an orthographic transcription will be given). The conventions used for the annotation are those defined within the SpeechDat project.

Ilustracja 1: Polycost - informacje z katalogu ELRA

1.2. Baza *GlobalPhone - GlobalPhone Polish*

Korpus *GlobalPhone Polish* stanowi część wielojęzykowego korpusu nagrań *GlobalPhone*. Korpus jako całość został zaprojektowany do wykorzystania w procesie tworzenia i ewaluacji systemów identyfikacji mówcy, rozpoznawania / identyfikacji języka i rozpoznawania mowy.

- **dostępność**: katalog ELRA, od 600 do 3600 € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=1142

W każdym języku 100 mówców odczytuje ok. 100 wypowiedzi. Teksty wybrano z gazet dostępnych w internecie (teksty dostarczają ok. 65 tys. wyrazów). Treść tekstów: tematyka polityczna krajowa imiędzynarodowa oraz wiadomości gospodarcze. Nagrania w 16bit, 16kHz mono, z mikrofonu Sennheiser 440-6 (na takim samym sprzęcie dla wszystkich języków).

Transkrypcja nagrań w korpusie *GlobalPhone* została poddana walidacji i na tym etapie uzupełniono ją dodatkowymi znacznikami dla cech mowy spontanicznej, takich jak: jąkanie się, pomyłki (false starts), zjawiska niewerbalne np. śmiech, dźwięki zawahań / namysłu. Bazę uzupełniono o informacje dotyczące wieku, płci, wykonywanego zawodu. Jako całość, korpus *GlobalPhone* zawiera ponad 450 godzin mowy zrealizowanej przez 1900 natywnych mówców dorosłych. Oprócz polskiego korpus *GlobalPhone* zawiera nagrania mowy czytanej dla 19 języków: arabski, bułgarski, chiński-mandaryński, chiński (Szanghaj), chorwacki, czeski, francuski, niemiecki, japoński, koreański, portugalski (brazylijski), rosyjski, hiszpański (płd. amerykański), szwedzki, tajski, tamilski, turecki, wietnamski (więcej, por. Schultz 2002)

The Polish part of GlobalPhone was collected from altogether 102 native speakers in Poland, of which 48 speakers were female and 54 speakers were male. The majority of speakers are between 20 and 39 years old, the age distribution ranges from 18 to 65 years. Most of the speakers are non-smokers in good health conditions. Each speaker read on average about 100 utterances from newspaper articles, in total we recorded 10130 utterances. The speech was recorded using a close-talking microphone Sennheiser HM420 in a push-to-talk scenario. All data were recorded at 16kHz and 16bit resolution in PCM format. The data collection took place in small and large rooms, about half of the recordings took place under very quiet noise conditions, the other half with moderate background noise. Information on recording place and environmental noise conditions are provided in a separate speaker session file for each speaker. The text data used for recording mainly came from the news posted in an online edition of a national Polish newspaper Dziennik Polski, (http://www.dziennik.krakow.pl/). We followed the standard GlobalPhone protocols and focused on national and international politics and economics news (see [SCHULTZ 2002]). In sum, 10130 utterances were spoken. The transcriptions are provided in Polish script in UTF-8 encoding and are also mapped to Roman script (Ascii). The Polish data are organized in a training set of 82 speakers, a development set of 10 speakers, and an evaluation set of another 10 speakers.

Ilustracja 2: Informacje na temat zawartości bazy GlobalPhone Polish z katalogu ELRA

1.2. Korpus *Yoho*

Mowa czytana, język: amerykański angielski, mówcy nagrywani kilkukrotnie na przestrzeni 3 miesięcy, jeden kanał, nagrania w typowym środowisku biurowym, krótkie frazy liczbowe.

- **dostępność**: katalog LDC, 150$[5]
- strona w katalogu: www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S16

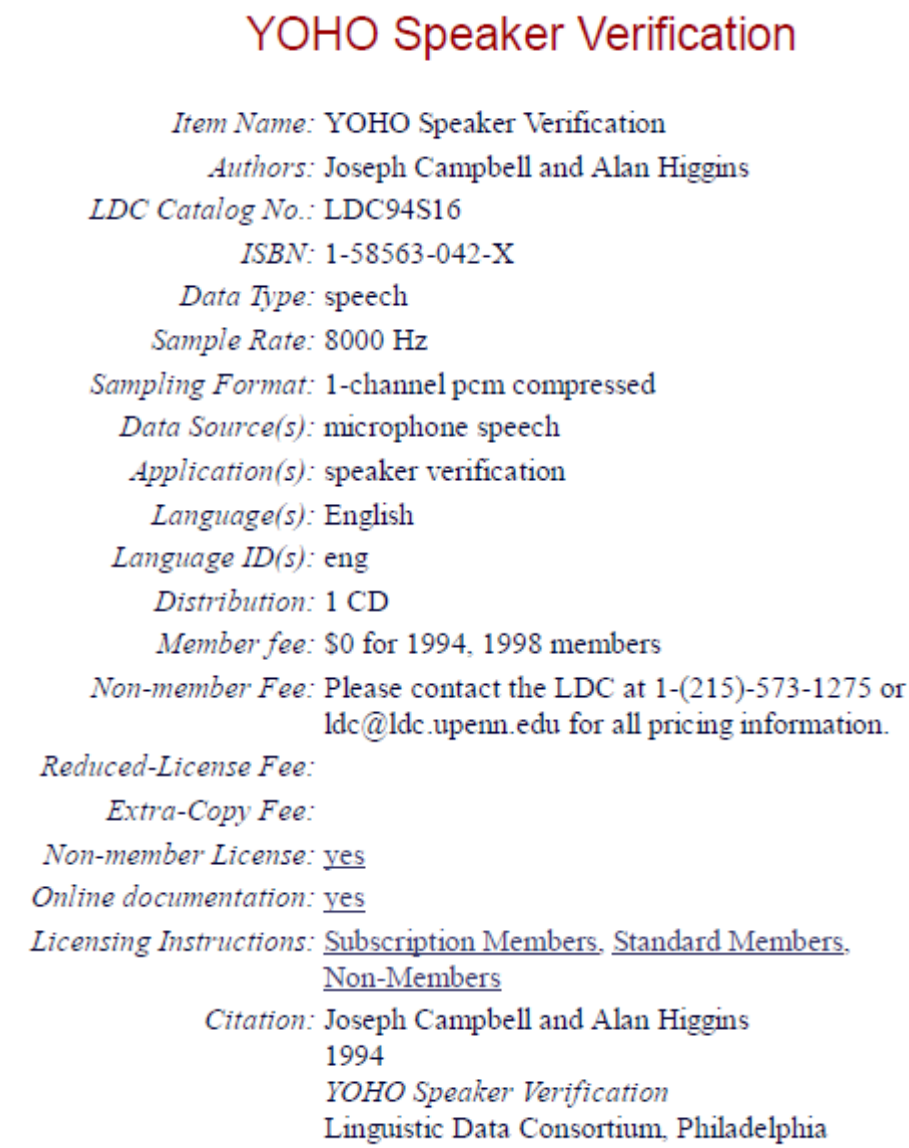
Korpus YOHO został zaprojektowany w celu wykorzystania w badaniach dotyczących autentyfikacji mówcy, 'text-dependent', np. na potrzeby systemów zabezpieczeń. Dane zebrano w 1989 roku w ramach umowy z rządem USA. Dane udostępniane obecnie w katalogu LDC zostały poddane anonimizacji, a część zawartości pierwotnego korpusu nie jest udostępniana publicznie (prawa dostępu zachował rząd USA).

Korpus YOHO zawiera:

- Frazy nagrane z przeznaczeniem do wykorzystania w szyfrach"Combination lock" (np. 36-24-36)
- Dane zebrane w okresie 3 miesięcy w rzeczywistym środowisku biurowym
- 4 'enrollment sessions' na mówcę, każda sesja - 24 frazy
- 10 sesji testowych na mówcę, każda sesja - 4 frazy
- Częstotliwość próbkowania 8kHz, analog bandwidth - 3.8 kHz
- 1.5 gigabajtów danych

Liczba prób jest wg autorów wystarczająca na potrzeby przeprowadzenia wiarygodnych testów ewaluacyjnych weryfikacji mówców. W każdej sesji mówca otrzymywał jako prompt kilka fraz do odczytania na głos; każda fraza to sekwencja liczb (np. "35-72-41", wymówione jako "thirty-five seventy-two forty-one"). Pierwsze cztery sesje dla każdego mówcy stanowiły 'enrollment sessions', składające się z 24 fraz. Wszystkie pozostałe sesje stanowiły 'verification trials', składające się z 4 fraz każda. Łącznie baza zawiera 552 enrollment sessions i 1,380 trial sessions. Odstęp między sesjami nagraniowymi to 3 dni.

5 Ceny z LDC podawane na dzień 20 września 2011. Po 29 września zamiast cen w katalogu pojawia się informacja: "Please contact the LDC at 1-(215)-573-1275 or ldc@ldc.upenn.edu for all pricing information." dla członków LDC z niektórych lat zasoby dostępne są bezpłatnie.



YOHO Speaker Verification

Item Name: YOHO Speaker Verification
Authors: Joseph Campbell and Alan Higgins
LDC Catalog No.: LDC94S16
ISBN: 1-58563-042-X
Data Type: speech
Sample Rate: 8000 Hz
Sampling Format: 1-channel pcm compressed
Data Source(s): microphone speech
Application(s): speaker verification
Language(s): English
Language ID(s): eng
Distribution: 1 CD
Member fee: $0 for 1994, 1998 members
Non-member Fee: Please contact the LDC at 1-(215)-573-1275 or ldc@ldc.upenn.edu for all pricing information.
Reduced-License Fee:
Extra-Copy Fee:
Non-member License: yes
Online documentation: yes
Licensing Instructions: Subscription Members, Standard Members, Non-Members
Citation: Joseph Campbell and Alan Higgins
1994
YOHO Speaker Verification
Linguistic Data Consortium, Philadelphia

Ilustracja 3: Informacje z katalogu LDC nt. korpusu Yoho

1.3. Korpus *KING-92*

Mowa czytana, język: amerykański angielski, 51 mówców (uwaga!: wyłącznie głosy męskie), jeden kanał, nagrania w środowisku biurowym oraz z telefonu, wypowiedzi pół-spontaniczne (rozwiązywanie zadań, opisywanie obrazków itp.)

- **dostępność**: katalog LDC, 1250$
- strona w katalogu: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95S22

Dane zebrano w ITT w 1987 roku w ramach umowy z rządem USA. Baza obecnie dostępna w katalogu LDC oprata jest na danych przetworzonych w 1992 roku (szczegóły w katalogu LDC, stąd dodany człon w nazwie KING-92, wcześniej udostępniano bazę z 1987 roku pod nazwą KING, wersja '92 jest poprawiona). Baza zawiera nagrania 51 głosów męskich w dwóch wersjach, różniących się charakterstyką kanału: 1) telefonicznego 2) nagranego za pomocą wysokiej jakości mikrofonu. Mówców podzielono na 2 grupy: 25 (z laboratorium w Nutley) i 26 osób (z San Diego). Na każdego mówcę i kanał przypada 10 plików, zawierających 30-60 sekund nagrania. Odstęp między sesjami nagraniowymi waha się od tygodnia do miesiąca. Transkrypty zawierają ok 54 tys tokenów wyrazowych.

Bazę KING-92 zaprojektowano przede wszystkim na potrzeby eksperymentów 'closed set' dla identyfikacji i weryfikacji mówców 'text-independent'. Zadania dialogowe polegały na rozwiązywaniu problemów, opisie obrazka, interpretacji rysunku itp. Podstawowe wykorzystanie korpusu to weryfikacja systemów dla telefonii, aczkolwiek należy zwrócić uwagę na to, że nagrania z tej bazy nie symulują prawdziwych rozmów telefonicznych, ze względu na jednostronny format nagrań. Dane można wykorzystać dzieląc każdą z 10 sesji na mówcę na różne podzbiory uczące i testowe, z możliwością 'multiple test sets'. Można z ich pomocą badać wpływ uczenia na jakość uzyskanych wyników lub zmienność jakości wyników przy różnych próbach testowych przy określonym zbiorze uczącym.

Obecnie dostępna jest wyłącznie ortograficzna transkrypcja danych w bazie, bez segmentacji (w starszej wersji udostępniano bardziej szczegółową transkr. ale w związku z błędami zrezygnowano z niej w aktualnej wersji bazy).

KING Speaker Verification

Item Name:	KING Speaker Verification
Authors:	Dr. Alan Higgins and Dave Vermilyea
LDC Catalog No.:	LDC95S22
ISBN:	1-58563-050-0
Data Type:	speech
Sample Rate:	8000 Hz
Sampling Format:	1-channel pcm compressed
Data Source(s):	telephone conversations
Application(s):	speaker verification
Language(s):	English
Language ID(s):	ENG
Distribution:	1 CD
Member fee:	$0 for 1995, 1998 members
Non-member Fee:	Please contact the LDC at 1-(215)-573-1275 or ldc@ldc.upenn.edu for all pricing information.
Reduced-License Fee:	
Extra-Copy Fee:	
Non-member License:	yes
Readme File:	yes
Online documentation:	yes
Licensing Instructions:	Subscription Members, Standard Members, Non-Members
Citation:	Dr. Alan Higgins and Dave Vermilyea 1995 *KING Speaker Verification* Linguistic Data Consortium, Philadelphia

Ilustracja 4: Informacje z katalogu LDC nt. korpusu King-92

1.4. *Speaker Recognition Corpus* ver. 1.1

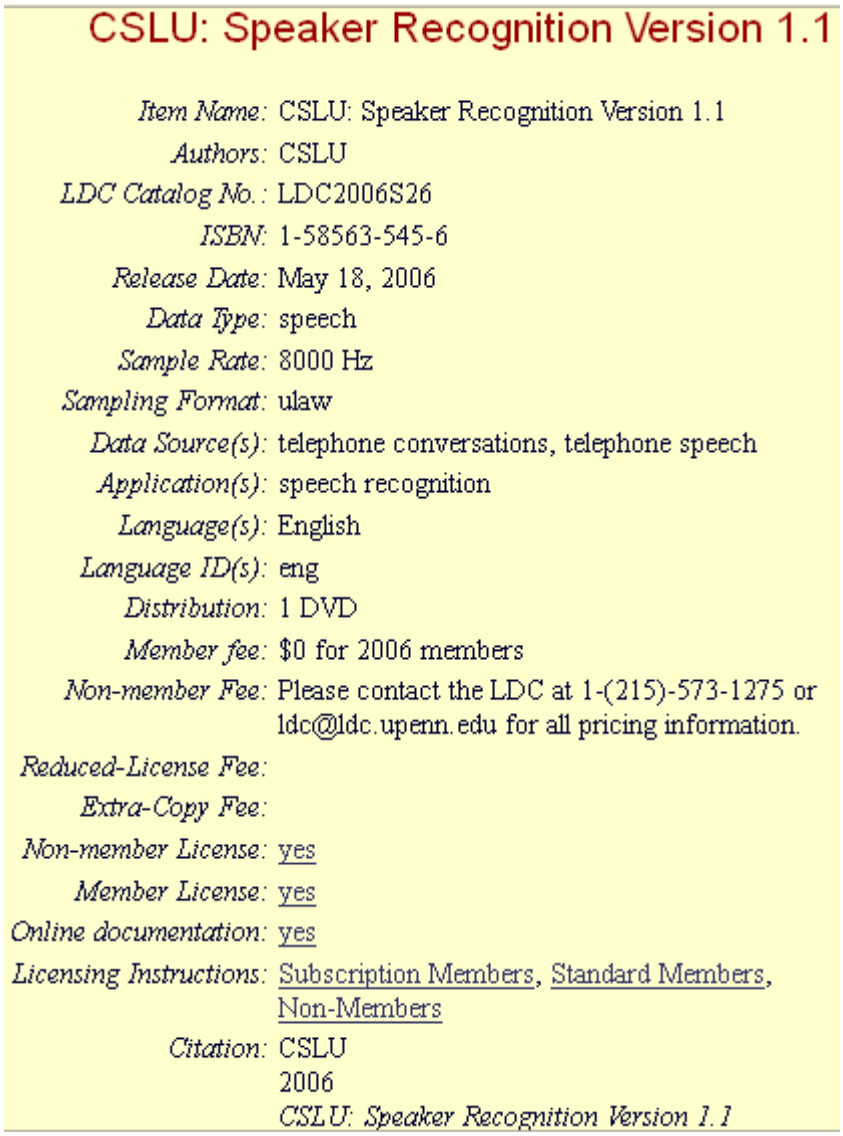
Speaker Recognition Corpus zawiera nagrania telefoniczne dla 91 głosów. Dla każdego mówcy zarejestrowano 12 sesji nagraniowych w okresie 2 lat.

- **dostępność**: katalog LDC
- strona w katalogu: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006S26

Podczas nagrań, każdy mówca wykonywał rejestrowaną rozmowę telefoniczną i odpowiadał na pytania typu: 'jaki jest twój kolor oczu' lub rejestrował opis na zadany temat np.'opisz typowy dzień z twojego życia'. Dla większości wypowiedzi dostępna jest transkrypcja na poziomie wyrazów (bez segmentacji). Podczas swoich dwunastu sesji nagraniowych każdy z mówców za każdym razem nagrywał odpowiedzi na te same pytania / tematy - bodźce.

Niektóre z sesji nagraniowych były rejestrowane w odstępach kilkudniowych, inne w dłużych odstępach, np. kilkutygodniowych. Uczestnicy nagrań postępowali według następującej procedury wykonywania rejestrowanych rozmów telefonicznych: podczas pierwszego miesiąca dzwonili dwa razy w tygodniu, w drugim i trzecim tygodniu nie wykonywano żadnych nagrań. W czwartym miesiącu - jeden telefon. Żadnych nagrań w piątym i szóstym miesiącu. Następnie ten sam harmonogram powtarzano trzykrotnie dla każdego z uczestników.

Aby zrównoważyć problemy związane z koordynacją prac przy tworzeniu bazy, np. z przypominaniem mówcom o terminach nagrań oraz z kumulacją większej ilości nagrań w systemie w tym samym czasie, grupę osób nagrywanych podzielono na 12 grup. Każda z grup rozpoczynała dwuletni okres nagrań w kolejnych miesiącach. Pierwsza grupa rozpoczęła we wrześniu 1996, druga w październiku tego samego roku itd.



CSLU: Speaker Recognition Version 1.1

Item Name:	CSLU: Speaker Recognition Version 1.1
Authors:	CSLU
LDC Catalog No.:	LDC2006S26
ISBN:	1-58563-545-6
Release Date:	May 18, 2006
Data Type:	speech
Sample Rate:	8000 Hz
Sampling Format:	ulaw
Data Source(s):	telephone conversations, telephone speech
Application(s):	speech recognition
Language(s):	English
Language ID(s):	eng
Distribution:	1 DVD
Member fee:	$0 for 2006 members
Non-member Fee:	Please contact the LDC at 1-(215)-573-1275 or ldc@ldc.upenn.edu for all pricing information.
Reduced-License Fee:	
Extra-Copy Fee:	
Non-member License:	yes
Member License:	yes
Online documentation:	yes
Licensing Instructions:	Subscription Members, Standard Members, Non-Members
Citation:	CSLU 2006 *CSLU: Speaker Recognition Version 1.1*

Ilustracja 5: Informacje z katalogu LDC nt korpusu Speaker Recognition

1.5. Korpus *Tactical Speaker Identification (TSID)*

Korpus zawiera nagrania 35 mówców (4kobiety, 31 mężczyzn), używających różnych nadajników i odbiorników radiowych.

- **dostępność**: katalog LDC
- strona w katalogu: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC99S83

Nagrania przeprowadzono dzieląc mówców na 7 grup po 5 osób, każda z osób wykonała następujące zadania:

- odczytanie listy zdań z korpusu TIMIT

- odczytanie listy ciągów liczb

- wytłumaczyć drogę z jednego punktu do drugiego na mapie (własnymi słowami)

- map-task - nagrywano obie osoby uczestniczące w zadaniu (ten etap dotyczy tylko 3 par mówców, dane zarejestrowano również wykorzystując różnego rodzaju ustawienia urządzeń rejestrujących).

Tactical Speaker Identification Speech Corpus (TSID)

Item Name: Tactical Speaker Identification Speech Corpus (TSID)
Authors: David Graff, Douglas Reynolds and Gerald C. O' Leary
LDC Catalog No.: LDC99S83
ISBN: 1-58563-154-X
Data Type: speech
Data Source(s): microphone speech
Application(s): speaker identification, speech recognition
Language(s): English
Language ID(s): eng
Distribution: 2 DVD
Member fee: $0 for 1999 members
Non-member Fee: Please contact the LDC at 1-(215)-573-1275 or ldc@ldc.upenn.edu for all pricing information.
Reduced-License Fee:
Extra-Copy Fee:
Non-member License: yes
Online documentation: yes
Licensing Instructions: Subscription Members, Standard Members, Non-Members
Citation: David Graff, Douglas Reynolds and Gerald C. O' Leary 1999 *Tactical Speaker Identification Speech Corpus (TSID)* Linguistic Data Consortium, Philadelphia

Ilustracja 6: Informacje z katalogu LDC nt korpusu Tactical Speaker Identification Speech Corpus (TSID)

1.6. Bazy zgodne z formatem SpeechDat dla języka angielskiego

Wszystkie trzy poniższe bazy SpeechDat dla j. angielskiego posiadają certyfikat walidacji SPEX na zgodność z formatem SpeechDat i na specyfikację zawartości.

SpeechDat(M) Polyphone database

Baza zawiera nagrania 1,000 mówców, nagranych za pośrednictwem linii telefonicznej. Język angielski, mowa czytana. Korpus jest podzielony na dwa podzbiory: część bazy znana jako **DB1** (sekwencje liczb, różnego rodzaju komendy, daty, godziny, kwoty pieniędzy itp.) (dwie płyty CD) oraz część zawierająca zdania bogate fonetycznie (jedna płyta CD) znana jako **DB2**. Częstotliwość próbkowania 8 KHz, 8bit. Dostępny jest leksykon wymowy w tej bazie wraz z transkrypcją w SAMPA.

Baza **Polyphone DB1**:

- **dostępność**: katalog ELRA, 6 tys. - 20 tys. € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=41

Baza **Polyphone DB2**:

- **dostępność**: katalog ELRA, 11 tys. - 20 tys. € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=42

British English SpeechDat(II) SDB-2400

Baza projektowana na potrzeby budowy i testowania systemów identyfikacji i weryfikacji mówcy. Zawiera nagrania 120 mówców, z których każdy wypowiada 22 jednostki po 20 razy. Dane zebrano z nagrań z telefonów stacjonarnych i komórkowych w warunkach cichych i z zakłóceniami.

- **dostępność**: katalog ELRA, między 39-47 tys. € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=770

Bazę podzielono na 8 płyt CD, nagrania 8-bit 8 kHz. Każda wypowiedź jest zapisana jako osobny plik dla każdego pliku jest dostępny plik ASCII SAM dostarczający informacji dotyczących nagrania i mówcy zgodnych z formatem SpeechDat. Dostępny jest też leksykon wymowy wraz z transkrypcją w SAMPA.

Each speaker uttered the following items:

* 1 sequence of 10 isolated digits
* 2 connected digits (1 credit card number -16 digits, 1 PIN code -6 digits)
* 2 spelled words (1 fixed "forename surname", 2 "names/words")
* 1 fixed "forename surname"
* 2 "forename surname" out of a set of 10
* 2 application words
* 10 phonetically rich sentences

The following age distribution has been obtained: 7 speakers are under 16, 41 speakers are between 16 and 30, 33 speakers are between 31 and 45, 32 speakers are between 46 and 60, and the age of 7 speakers is unknown.

Ilustracja 7: Szczegóły zawartości bazy British English SpeechDat(II) SDB-2400

2. Inne korpusy dla języka polskiego

2.1. Polish SpeechDat(E) Database

Korpus *Polish SpeechDat(E) Database* (*Eastern European Speech Databases for Creation of Voice Driven Teleservices*) zawiera nagrania 1000 mówców (488 mężczyzn, 512 kobiet) zarejestrowane za pomocą polskich stacjonarnych linii telefonicznych.

- **dostępność**: katalog ELRA, między 12-16 tys. € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=580

Corpus contents:
- 6 application words;
- 1 sequence of 10 isolated digits;
- 4 connected digits: 1 sheet number (5 digits), 1 telephone number (8-11 digits), 1 credit card number (15-16 digits), 1 PIN code (6 digits);
- 3 dates: 1 spontaneous date (birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 spotting phrase using an application word (embedded);
- 1 isolated digit;
- 3 spelled-out words (letter sequences): 1 spelling of surname; 1 spelling of directory assistance city name; 1 real/artificial name for coverage;
- 2 currency money amounts: 1 Polish money amount, 1 International money amount (USD, EURO)
- 1 natural number;
- 6 directory assistance names: 1 surname (out of 500); 1 city of birth / growing up (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 "forename surname" (set of 150), 1 "surname" (set of 150)
- 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question;
- 12 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

Ilustracja 8: Zawartość sesji nagraniowej w bazie Polish SpeechDat(E) Database

Baza *Polish SpeechDat(E)* jest podzielona na 5 płyt CD, każda zawiera 200 sesji nagraniowych. Pliki 8-bit, 8kHz, pliki nie są skompresowane zgodnie ze specyfikacją SpeechDat(E). Każda

wypowiedź jest zapisana jako osobny plik dla każdego pliku jest dostępny plik ASCII SAM. Baza posiada certyfikat walidacji SPEX na zgodność z formatem SpeechDat i na specyfikację zawartości.

2.3. Polish Speecon database

Baza SpeeCon dla języka polskiego - struktura scenariusza nagraniowego wg standardów Speecon, (uzyskanie zgodności z tym formatem planowane jest dla Jurisdic-t).

- **dostępność**: katalog ELRA, między 50-75 tys. € zależnie od zastosowań i członkostwa
- strona w katalogu: http://catalog.elra.info/product_info.php?products_id=801

Baza *Polish SpeeCon j*est podzielona na 2 podzbiory:

1) Pierwszy zbiór obejmuje nagrania 550 głosów dorosłych (286 mężczyzn, 264 kobiet), nagranych z użyciem czterech kanałów mikrofonowych w czterech różnych środowiskach zewnętrznych (biuro, "rozrywka", samochód, miejsce publiczne).

2) Drugi zbiór to nagrania 50 głosów dzieci (25 chłopców, 25 dziewczynek), nagrane z użyciem czterech kanałów w jednym środowisku zewnętrznym (pokój dziecięcy).

Wiek mówców w korpusie:

Dorośli: 286 mówców między 15 a 30 rokiem życia, 165 mówców między 31 a 45 r.ż, 79 mówców między 46 a 60 r.ż., 20 mówców powyżej 60 roku życia

Dzieci: 23 mówców między 8 a 10 r.ż., 27 mówców między 11 a 14 rokiem życia.

Dane *Polish SpeeCon* udostępnia się na 26 płytach DVD (część pierwsza, nagrania) and trzech płytach DVD z dokumentacją. Nagrania 16 kHz, 16 bit. Każda wypowiedź jest zapisana jako osobny plik dla każdego pliku jest dostępny plik ASCII SAM. Dostępny jest leksykon wymowy z transkrypcją w SAMPA. Baza posiada certyfikat walidacji SPEX na zgodność z formatem SpeechDat i na specyfikację zawartości.

Calibration data:
6 noise recordings
The "silence word" recording
Free spontaneous items (adults only):
3 minutes (session time) of free spontaneous, rich context items (story telling) (an open number of spontaneous topics out of a set of 30 topics)
17 Elicited spontaneous items (adults only):
3 dates, 2 times, 3 proper names, 2 city name, 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language
Read speech:
30 phonetically rich sentences uttered by adults and 60 uttered by children
5 phonetically rich words (adults only)
4 isolated digits
1 isolated digit sequence
4 connected digit sequences
1 telephone number
3 natural numbers
1 money amount
2 time phrases (T1 : analogue, T2 : digital)
3 dates (D1 : analogue, D2 : relative and general date, D3 : digital)
3 letter sequences
1 proper name
2 city or street names
2 questions
5 special keyboard characters (2 by children)
1 Web address
1 email address
208 application specific words and phrases per session (adults)
74 toy commands, 34 general commands and 14 phone commands (children)

Ilustracja 9: Polish Speecon - struktura scenariusza nagraniowego

2.4. Bazy mowy (pół)spontanicznej dla j. polskiego, korpusy multimodalne

- Baza **PoInt** (M. Karpiński) - maptask, dialogi kontrolowane (odgadnięcie postaci ze zdjęcia, opis zdarzenia z dzieciństwa, zabawnego komiksu) + czytane, 50 mówców w monologach + dialogi, całość obejmuje około 30 godzin nagrań, udostępniana tylko część (anotacja w formacie programu Praat)
- **DiaGest** (M. Karpiński) - video i audio, dwie kamery - 10 dzieci, 10 dorosłych, video, zorientowane na gestykulację, akty dialogowe, anotowane (format programu ELAN)
- **Diagest2** (M. Karpiński) - video i audio - zadanie 'origami' (wykonywanie zadanego elementu z papieru, szczegóły zob. część obecnego raportu dotycząca cech paralingwistycznych) anotowane (format programu ELAN)
- **Luna** - nagrania z informacji telefonicznej Zakładu Transportu Miejskiego w Warszawie (anotacja nawiązująca do formatu SpeeCon, niektóre aspekty uszczegółowione),
  - 500 dialogów człowiek-maszyna,
  - łączny czas wypowiedzi 670 minut,
  - liczba wypowiedzi 12 788; 500 dialogów człowiek-człowiek
  - 100 dialogów dla systemu opłat (14 mówców)
    - http://www.ist-luna.eu/
    - http://nlp.ipipan.waw.pl/NLP-SEMINAR/071112.pdf

3. Bazy danych wykorzystywane w badaniu emocji

Najnowsze bazy danych dla mowy emocjonalnej czy ekspresywnej są często multimodalne. Zakłada się, że w badaniach cech paralingwistycznych / pozajęzykowych potrzebny jest kontekst szerszy, aniżeli wyłącznie sygnał dźwiękowy. Zarówno uzyskanie nagrań dostarczających informacje o autentycznych emocjach mówcy, jak i opis oraz interpretacja materiału zawierającego nagrania mowy nacechowanej emocjonalnie stanowi bardzo duże wyzwanie. Mimo, że badania w tej dziedzinie trwają od lat (por. część obecnego raportu dotycząca analizy i modelowania emocji), wyniki nadal nie są zadowalające i w niewystarczającym stopniu spełniają wymogi stawiane w zastosowaniach praktycznych.

Optymalnie, gdy udaje się uzyskać autentyczne dane dźwiękowe, takie jak nagrania ze źródeł policyjnych (autentyczne emocje, telefony alarmowe: wysokie nacechowanie emocjonalne, sytuacje ekstremalne, nagrania rozpraw sądowych, przesłuchań - wysoki poziom stresu). Są one jednak trudno dostępne ze względów prawnych i etycznych. Ponadto zakres ich zastosowań jest ograniczony, ze względu na określoną specyfikę zawartości emocjonalnej, są to nagrania często nie nadające się do analizy emocji w standardowej sytuacji komunikacyjnej.

W niektórych zastosowaniach dobrym rozwiązaniem są nagrania z różnego rodzaju central telefonicznych (nagrania mowy spontanicznej o słabszym/znikomym nacechowaniu emocjonalnym) - mogą być łatwiejsze w pozyskaniu, ale nadal istnieją problemy podobnie jak w przypadku baz policyjnych, np. dane osobowe, ponadto także problem zawartości i użyteczności takich nagrań w badaniach nad mową emocjonalną / ekspresywną w naturalnych sytuacjach komunikacyjnych.

Spotyka się także korpusy zawierające nagrania z telewizji lub radia, dla różnych typów sytuacji komunikacyjnych (wywiady, debaty, reality show). Tutaj także za każdym razem istotna jest kwestia prawna, ale nie tylko. Zawsze należy mieć na uwadze wpływ szerszego kontekstu (kamery, świadomość bycia oglądanym, ocenianym itp.), również nietypowego dla warunków naturalnych.

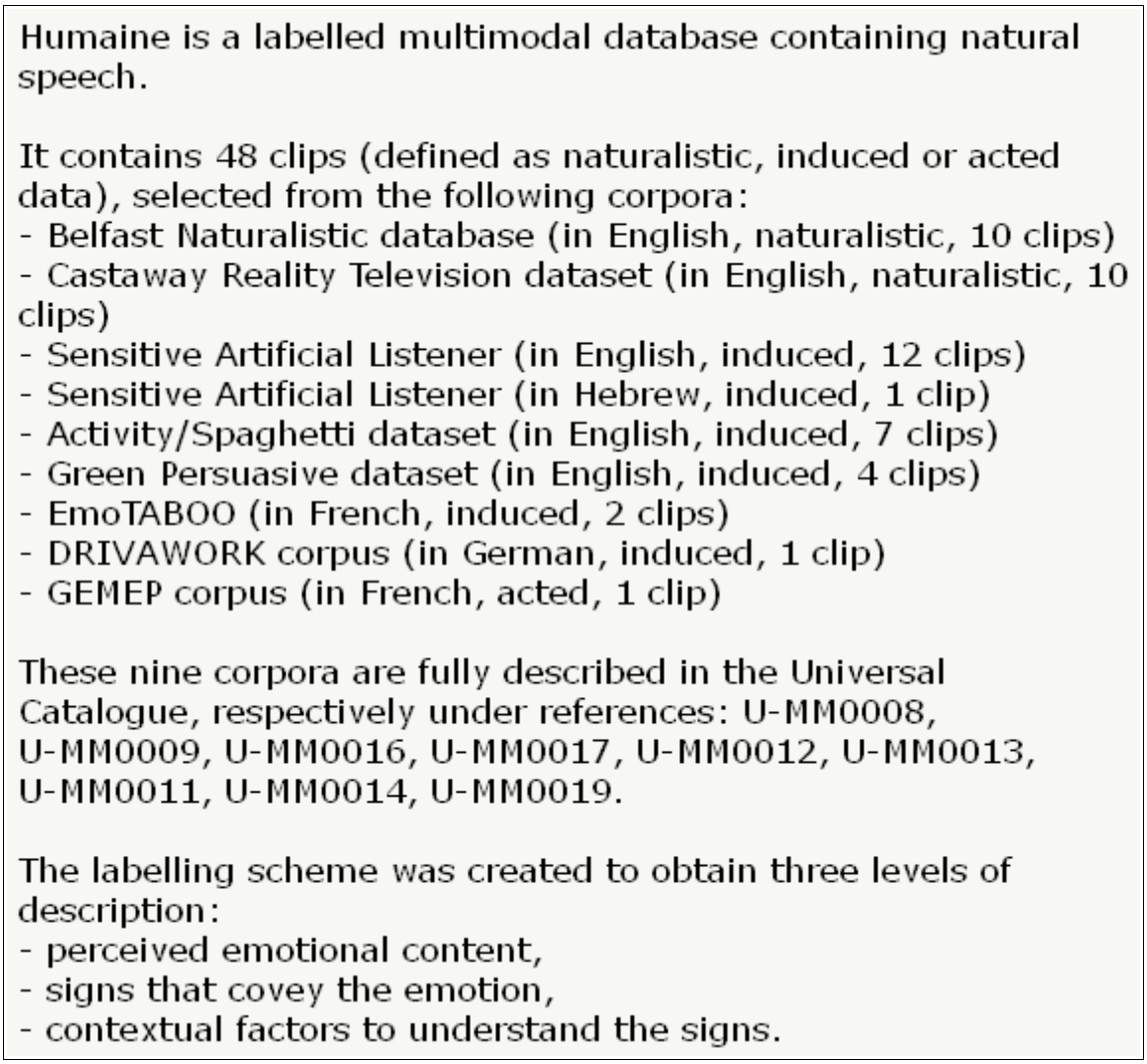
3.1. Projekt i baza danych *Humaine*

Jednym z większych projektów badawczych ukierunkowanych na badanie emocji w mowie i ogólnie w komunikacji człowiek-człowiek, a także człowiek-komputer, także na potrzeby zastosowań praktycznych, jest projekt *Humaine*. Jednym z jego wyników jest korpus *Humaine*, zawierający dane nowe lub pozyskane z już istniejących /powstających równolegle baz 'emocjonalnych'.

Baza źródłowa	język	charakter materiału nagraniowego	ilość danych	dostępność
Belfast Naturalistic Database (Douglas-Cowie et al. 2003)	angielski	emocje naturalne/ wywoływane; emocje raczej umiarkowane, różnego rodzaju, nagrania AV (audio-video) z telewizji i wywiady	298 nagrań, 31 mężczyzn, 94 kobiet, każdy po 2 sekwencje do 60 sek (wypowiedź neutralna i nacechowana): 239 klipów, w tym 209 z telewizji, 30 z wywiadów + percepcyjne etykiety emocji z Feeltrace	udostępniono wybrane 30 sekwencji. kontakt: c.cox@qub.ac.uk
Castaway Reality Television Dataset (Douglas-Cowie et al. 2007)	angielski	naturalne; w większości emocje umiarkowane, niektóre bardziej intensywne, nagrania; nagrania na wyspie, grupa osób bierze udział w różnego rodzaju działaniach (gaszenie ognia, reakcja na węża)	10 taśm, 30 minut każda	wg publikacji zwolnione z praw autorskich, w internecie - nie znalazłam, trzeba by prawdopodobnie skontaktować się z autorami
Sensitive Artificial Listener SAL: - Belfast (ang) - TelAviv (hebr) (Douglas-Cowie et al. 2007)	angielski / hebrajski	wywoływane, umiarkowane emocje; komunikacja człowiek-komputer przez interfejs SAL: 4 'osobowości', które mają sprowokować stany emocjonalne u mówców	4 sesje w j. angielskim, ok 20 minut każda, interfejs przetłumaczony następnie na hebrajski (przy tłumaczeniu dostosowano także kulturowe aspekty zachowań postaci SAL); http://wwwhome.ewi.utwente.nl/~hofs/sal/index.html	dostępne na potrzeby badań naukowych, kontakt: c.cox@qub.ac.uk
Activity Data/Spaghetti Data (Douglas-Cowie et al. 2007)	angielski	wywoływane; nagrania AV uzyskane za pomocą 2 metod, w czasie: 1) aktywności zewn., np. wyścigi rowerów górskich (pozytywne i negatywne emocje, stosunkowo wysokie pobudzenie), 2) wkładania ręki do pudełka z zaskakującą/dziwną zawartością (zaskoczenie, szok, obrzydzenie...)	ok. 60 mówców	wg publikacji dostępne na potrzeby badań naukowych; w internecie - nie znalazłam, trzeba by prawdopodobnie skontaktować się z autorami
Green Persuasive Dataset	angielski	wywoływane; jeden z rozmówców o proekologicznych poglądach przekonuje drugiego do bardziej ekologicznego stylu życia, później ta przekonywana osoba opowiada na ile przekonana się czuje	16 mówców, 8 sesji po ok. 30 minut, plus nagrania komentarzy osób przekonywanych	dostępne za darmo na potrzeby niekomercyjnych badań: http://sspnet.eu/2009/12/green-persuasive-database/
EmoTABOO (Zara et al. 2007)	francuski	wywoływane; mówcy podczas gry pt. Taboo, jedna osoba tłumaczy drugiej znaczenie słowa/pojęcia, używając gestów i ruchów ciała	różne emocje, w tym zakłopotanie, rozbawienie	należy się skontaktować z zespołem LIMSI: http://www.limsi.fr/
EmoTV Database (Devillers et al. 2006)	francuski	wywoływane; w większości emocje umiarkowane, niektóre bardziej intensywne, nagrania AV wywiadów telewizyjnych (zarówno studyjne jak i w plenerze)	51 klipów dla 48 mówców (4-43 sek. nagrania nacechowanego emocjonalnie na mówcę)	dostęp ograniczony, część danych może być udostępniana po indywidualnych negocjacjach.
DRIVAWORK (Driving under Varying Workload) (Honig 2006)		bardziej nastawione na stany związane **ze stresem** niż z innymi emocjami, nagrania podczas symulacji kierowania samochodem, różne rodzaje zadań: kierowanie w stanie rozluźnienia oraz z dodatkowym zadaniem do rozwiązania (np. matematycznym); nagrania video oraz rejestracja objawów fizjologoicznych (elektrokardiografia, temperatura skóry, oddychanie i in.)	24 uczestników, 15 h fizjologicznych nagrań (relaxed vs. stressed)	należy skontaktować się z zespołem Ernlangen http://www5.informatik.uni-erlangen.de/en/our-team/hoenig-florian
GEMEP Corpus (Geneva Multimodal Emotion Portrayal)	francuski	odgrywane; aktorzy odgrywają zadane emocje = 'portrety emocjonalne'	10 aktorów, 18 'emocji' - łącznie 1260 'portetów emocji' (*emotion portayals*)	

*Tabela 2. Baza '**Humaine**' - informacje o źródłowych bazach składowych*

W tabeli wyżej podsumowano informacje na temat baz składowych, z których wybrano ostateczne składowe bazy *Humaine*. Według deklaracji na stronach projektu Humaine http://emotion-research.net/ część zasobów jest dostępna nieodpłatnie. Poniżej natomiast zamieszczono informacje z katalogu ELRA Universal o składzie ostatecznie wybranych składowych w bazie Humaine.


Humaine is a labelled multimodal database containing natural speech.

It contains 48 clips (defined as naturalistic, induced or acted data), selected from the following corpora:
- Belfast Naturalistic database (in English, naturalistic, 10 clips)
- Castaway Reality Television dataset (in English, naturalistic, 10 clips)
- Sensitive Artificial Listener (in English, induced, 12 clips)
- Sensitive Artificial Listener (in Hebrew, induced, 1 clip)
- Activity/Spaghetti dataset (in English, induced, 7 clips)
- Green Persuasive dataset (in English, induced, 4 clips)
- EmoTABOO (in French, induced, 2 clips)
- DRIVAWORK corpus (in German, induced, 1 clip)
- GEMEP corpus (in French, acted, 1 clip)

These nine corpora are fully described in the Universal Catalogue, respectively under references: U-MM0008, U-MM0009, U-MM0016, U-MM0017, U-MM0012, U-MM0013, U-MM0011, U-MM0014, U-MM0019.

The labelling scheme was created to obtain three levels of description:
- perceived emotional content,
- signs that covey the emotion,
- contextual factors to understand the signs.

Ilustracja 10: Humaine' - składowe bazy wg katalogu ELRA-U-MM0020 (Universal Catalogue)

Efekty działań w ramach projektu *Humaine* nie ograniczają się wyłącznie do korpusu *Humaine*; na stronie http://emotion-research.net/ znajduje się ponadto wiele innych informacji na temat aktualnie tworzonych zasobów i narzędzi na potrzeby badań mowy nacechowanej emocjonalnie. Jednym z najbardziej interesujących projektów jest tworzenie bazy **CREST** (por. Campbell, 2000) (rejestracja danych: mówcy nagrywani za pomocą przenośnych zestawów

nagrywających, przez okres ok. 5 lat, w warunkach quasi-naturalnych, w codziennych sytuacjach, docelowo zaplanowana wielkość bazy 1000 godzin nagrań, trzy języki: angielski, japoński, chiński). **Dostępność** - nie znaleziono w katalogach, prawdopodobnie należałoby skontaktować się z autorami.

Rozpatrując wykorzystanie różnego rodzaju zewnętrzynych nagrań mowy emocjonalnej czy ekspresywnej w rozwoju aplikacji technologii mowy, należy również mieć na uwadze możliwe zależności sposobu wyrażania emocji i ekspresji głosowej od uwarunkowań językowych i kulturowych. Mimo wskazywanych w niektórych badaniach uniwersaliów dotyczących tych właśnie sfer (zob. część II obecnego raportu, dotycząca emocji), zalecana byłaby zatem ostrożność, jeśli idzie o to, na ile są lub nie są one zależne od języka, w którym formułowany jest komunikat (por. *The ways in which culture shapes emotional expression are likely to be complex*, S.R. Fussel w: 'The Verbal Communication of Emotion').

Literatura

Allwood J., Cerrato L., Dybkjaer L., Jokinen K., Navarretta C., Paggio P., *The MUMIN Multimodal Coding Scheme*, NorFA Yearbook 2005.

Biadsy, F., William Yang Wang, Andrew Rosenberg, Julia Hirschberg 2011. Intoxication Detection Using Phonetic, Phonotactic and Prosodic Cues. Proceedings of InterSpeech 2011, Florence.

Bjorn Schuller, Zixing Zhang, Felix Weninger, Gerhard Rigoll 2011. Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? Proceedings of Interspeech 2011, Florence, Italy.

Burkhardt, F.; M. Eckert, W. Johannsen, and J. Stegmann 2010. A Database of Age and Gender Annotated Telephone Speech, Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, 2010.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., & Auberge, V. (2006). *Emotional prosody - does culture make a difference?. Proc. Speech Prosody 2006.*

Campbell, N. 2007. Differences in the Speaking Styles of a Japanese Male According to Interlocutor; Showing the Effects of Affect in Conversational Speech. *Computational Linguistics and Chinese Language Processing.* vol. 12, No. 1, March 2007, pp. 1-16.

Campbell, N. Voice characteristics of spontaneous speech.

Campbell, N. (2002), "The Recording of Emotional Speech; JST/CREST database research", Proc. of LREC2002, Vol. 6, 2029-2032.

Carletta, J. (2007) Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation Journal 41(2): 181-190.

Charfuelan, M., Schroder, M. 2011. The Vocal Effort of Dominance in Scenario Meetings. Proceedings of InterSpeech 2011, Florence.

Cole, R., Mike Noel, Victoria Noel (1998) The CSLU Speaker Recognition Corpus. ICSLP, Sydney, Australia, Nov. 30, 1998.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. 2001. Emotion Recognition in Human-Computer Interaction". Signal Processing Magazine, vol. 18, str. 32-80.

Crystal, D.: The linguistic status of prosodic and paralinguistic features.

Davies, Alan, & Henry Widdowson. 1974. Reading and writing. Techniques in Applied Linguistics, ed. by J. Allen & S. Corder, 154-201. London: Oxford University Press.

Devillers, L., Vidrascu, L. 2006. Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs, in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, PA, pp. 801–804.

Emotion Classification of Infants' Cries Using Duration Ratios of Acoustic Segments *K. Kitahara, S. Michiwiki, M. Sato, S. Matsunaga, M. Yamashita, K. Shinohara; Nagasaki University, Japan Proceedings of Interspeech 2011, Florence, Italy.*

Eyben, F., W¨ollmer, M. and Schuller, B. 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, in Proc. ACII, Amsterdam, Netherlands, pp. 576–581.

Fanelli G., Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool, Member, IEEE (2010) A 3-D Audio-Visual Corpus of Affective Communication , [w:] IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 12, NO. 6, OCTOBER 2010 591

Frederike Gottsmann, Corinna Harwardt 2011. Investigating Robustness of Spectral Moments on Normal- and High-Effort Speech, Proceedings of Interspeech 2011, Florence, Italy.

Fujisaki, H., Hirose, K. 1993. Analysis and perception of intonation expressing paralinguistic information in Japanese. ESCA Workshop on Prosody, Lund.

Gobl, C. and A. Ní Chasaide 2003. The role of voice quality in communicating emotion, mood

and attitude, *Speech Communication,* vol. 40, pp. 189-212.
Grimm, M., Kroschel, K., Narayanan, S. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database, in Proc. of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, pp. 865–868.
Gussenhoven, C. 2004. The Phonology of Tone and Intonation, Cambridge: CUP.
Gut, U., Looks K., Thies A., Trippel T., Gibbon D., *CoGesT – Conversational Gesture Transcription System - Version 1.0*, Bielefeld.
Hagen, A. 1997. Linguistic Functions of Glotalizations and their Language Specific Use in English
Hennebert, J., H. Melin, D. Petrovska, D. Genoud (2000) POLYCOST: A telephone-speech database for speaker recognition, Speech Communication Volume 31, Issues 2-3, June 2000, Pages 265-270 and German. Ph.D. disertation, Friedrich-Alexander-UniversitäErlangen-Nürnberg and MIT.tst
Hewlett-Sanchez, M., D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, W. Jarrold 2011. Using Prosodic and Spectral Features in Detecting Depression in Elderly Males. Proceedings of Interspeech 2011, Florence.
Heylen, D., Anton Nijholt, Dennis Reidsma Determining what people feel and think when interacting with humans and machines Notes on corpus collection and annotation Human Media Interaction University of Twente, The Netherlands {heylen,anijholt,dennisr}@ewi.utwente.nl
Hönig F, Batliner A, Eskofier B, Nöth E (2008) Predicting Continuous Stress Ratings Of Multiple Labellers From Physiological Signals, w: BIOSIGNAL 2008 (Drivavork corpus)
Ishi, C., Ishiguro, H., Hagita, N. 2006. Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction, in Proc. of Speech Prosody 2006, Dresden, pp. 883–886.
Ishi, C.T., Hiroshi Ishiguro, Norihiro Hagita;Analysis of Acoustic-Prosodic Features Related to Paralinguistic Information Carried by Interjections in Dialogue Speech. Proceedings of Interspeech 2011, Florence, Italy.
Ivanov, A.V., *G. Riccardi, A.J. Sporka, J. Franc 2011.* Recognition of Personality Traits from Human Spoken Conversations. Proceedings of Interspeech 2011, Florence, Italy.
Jangwon Kim, Sungbok Lee, Shrikanth Narayanan 2011. An Exploratory Study of the Relations Between Perceived Emotion Strength and Articulatory Kinematics. Proceedings of InterSpeech 2011, Florence.
Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z., Szczyszek, M. 2008. Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych. Investigationes Linguisticae, vol. XVI.
Khiet P. Truong, Ronald Poppe, Iwan de Kok, Dirk Heylen; A Multimodal Analysis of Vocal and Visual Backchannels in Spontaneous Dialogs. Proceedings of Interspeech 2011, Florence, Italy.
Kipp M., Neff M., Albrecht I. 2007. *An Annotation Scheme for Conversational Gestures: How to economically capture timing and form,* [w:] *Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora*, Vol. 41, Nr. 4, Springer.
Kitazawa, S., Kobayashi, S., Matsunaga, T., Ichikawa, H. 1994. Acoustic Analysis and Transcription of Linguistic and Paralinguistic Features in Dialogue Speech. Studia Phonologica, vol. 28, pp. 11-23.
Kohler, K. 2003. Pragmatic and Attitudinal Meanings of Pitch Patterns in German Syntactically Marked Questions.
Lamel, L.F., Jean-Luc Gauvain, J.-L., 1993. Identifying Non-Linguistic Speech Features. Proc. EuroSpeech 1993, pp. 23-31.
Lausberg, H., Sloetjes, H. 2009. Coding gestural behavior with the NEUROGES–ELAN system. *Behavior Research Methods 2009, 41 (3), 841-849*
Liscombe, J.J. 2007. Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency. PhD Dissertation, Columbia University.
Longe, 1999. THE LINGUISTIC REALIZATION OF PARALINGUISTIC FEATURES IN ADMINISTRATIVE LANGUAGE. Studies in the Linguistic Sciences Volume 29, Number 1 (Spring 1999) .
Marasek, K. and Gubrynowicz, R. 2004. Multi-level Annotation in SpeeCon Polish Speech Database, Proceedings of IMTCI, str. 58-67.
Marciniak, M. (red.) 2010. *Anotowany korpus dialogów telefonicznych.* Warszawa: EXIT.
McCowan, I.M., J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J.

Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner.2007. The AMI Meeting Corpus. *The AMI Project Consortium, www.amiproject.org.*

Meghjani, M., Frank Ferrie and Gregory Dudek. 2009. Bimodal information analysis for emotion recognition. IEEE, Workshop on Applications of Computer Vision (WACV), Utah, USA. December 2009.

Montacie, C., Caraty, M.J. 2011.. Combining Multiple Phoneme-Based Classifiers with Audio Feature-Based Classifier for the Detection of Alcohol Intoxication.

Mori, H., Satake, T., Nakamura, M., Kasuya, H. 2011. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics, *Speech Communication*, Volume 53, Issue 1, str. 36-50.

Mori, Kasua, Nakamura 2010. Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies.

Mueller, C. 2007. Classifying speakers according to age and gender, in Speaker Classification II, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Mueller, Ed. Heidelberg - New York - Berlin: Springer, 2007, vol. 4343.

Murray, I. and Arnott, J. 1996. Synthesizing Emotions in Speech: Is it time to get excited? In Proceedings of 4th Int. Conf. Spoken Language Processing, pp. 1816-1819.

Neiberg, D., Petri Laukka, Hillary Anger Elfenbein 2011. Intra-, Inter-, and Cross-Cultural Classification of Vocal Affect. Proceedings of Interspeech 2011, Florence, Italy.

Nguyen, Q., Kipp, M. 2010. Annotation of Human Gesture using 3D Skeleton Controls. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, ELDA, Paris.

Noeth, W. (1990). Handbook of Semiotics. Indiana University Press, Bloomington & Indianapolis.

Nomoto, N., Masafumi Tamoto, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi, 2011. Anger Recognition in Spoken Dialog Using Linguistic and Para-Linguistic Information.

Oertel, C., Stefan Scherer, Nick Campbell. On the Use of Multimodal Cues for the Prediction of Degrees of Involvement in Spontaneous Conversation.

Pakhomov, S., Kotlyar, M. 2011. Prosodic Correlates of Individual Physiological Response to Stress. Proceedings of InterSpeech 2011, Florence.

Patel, S., Shrivastav, R. 2011. A Preliminary Model of Emotional Prosody Using Multidimensional Scaling. Proceedings of InterSpeech 2011, Florence.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W. 2005. The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. Speech Communication, 45, 90-95.

Rahman, T., Soroosh Mariooryad, Shalini Keshavamurthy, Gang Liu, John H.L. Hansen, Carlos Busso 2011. Detecting Sleepiness by Fusing Classifiers Trained with Novel Acoustic Features. Proceedings of InterSpeech 2011, Florence.

Roach, P. 2000. TECHNIQUES FOR THE PHONETIC DESCRIPTION OF EMOTIONAL SPEECH.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998). *Transcription of prosodic and paralinguistic features of emotional speech. Journal of the International Phonetic Association: 28*, 83–94.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92.

Schoetz, S., Mueller, Ch. 2007. A Study of Acoustic Correlates of Speaker Age. In *Speaker Classification*, Lecture Notes in Computer Science / Artificial Intelligence 4343. 2-9.

Schötz, S. 2002. Linguistic & Paralinguistic Phonetic Variation in Speaker Recognition & Text-to-Speech Synthesis (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.1401&rep=rep1&type=pdf)

Schröder, M., Elisabetta Bevacqua, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Valstar, Wöllmer (2009) A Demonstration of Audiovisual Sensitive Artificial Listeners

Schuller, B., Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, Vered Aharonson 2007. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. Proceedings of InterSpeech 2007.

Schuller, B., Rigoll, G., Lang, M. "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture". IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, no.1, pages I-577-80, May 2004.

Schuller, B., Steidl, S., Batliner, A. 2009. "The INTERSPEECH 2009 Emotion Challenge," in Proc. Interspeech, Brighton, UK, pp. 312–315.

Schuller, B.; R. Mueller, F. Eyben, J. Gast, B. H¨ornler, M. W¨ollmer, G. Rigoll, A. Hoethker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior, vol. 27, pp. 1760–1774, 2009.

Schuller, B.; R. Mueller, F. Eyben, J. Gast, B. Hoernler, M. Woellmer, G. Rigoll, A. Hoethker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior, vol. 27, pp. 1760–1774, 2009.

Schultz, T. (2002) GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University, Proceedings of the International Conference of Spoken Language Processing, ICSLP 2002, Denver, CO, September 2002.

Stadelmann, T. (2010) Voice Modeling Methods for Automatic Speaker Recognition, rozprawa doktorska: http://www.mathematik.uni-marburg.de/~stadelmann/download/papers/thesis.pdf

Trager, G.L. 1958. Paralanguage: A first approximation. Studies in Linguistics, 13, pp. 1-12.

Vlasenko, B., Dmytro Prylipko, David Philippou-Hubner, Andreas Wendemuth 2011. Vowels Formants Analysis Allows Straightforward Detection of High Arousal Acted and Spontaneous Emotions. Proceedings of Interspeech 2011, Florence, Italy.

Weninger, F., Schuller, B., Woellmer, M., Rigoll, G. 2011. Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory. Proceedings of ICASSP.

Yaeger-Dror, M. 2002. Register and prosodic variation, a cross language comparison, *Journal of Pragmatics,* 34 (2002), 1495–1536.

Yanushevskaya, I., C. Gobl, and A. Ní Chasaide 2008. Voice quality and loudness in affect perception, in *Speech Prosody 2008*, Campinas, Brazil.

Zara, A., V. Maffiolo, J.-C. Martin, and L. Devillers, "Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics," in Proc. ACII, 2007, pp. 464–475.

Zeng, Z., Pantic, M., Roisman, G., Huang, T. 2009. A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39-58, January, 2009.