Intonation processing for speech technology

Przetwarzanie intonacji na potrzeby technologii mowy

Grażyna Demenko

Institute of Linguistics, Department of Phonetics
Adam Mickiewicz University, Poznań
lin@amu.edu.pl

ABSTRACT

The paper presents problems arising in the analysis of intonation patterns in speech, their interpretation and usage for the needs of Speech Technology. Multiple functions of intonation are discussed from points of view of acoustic phonetics, linguistics and psychology. The results of a comprehensive, multilevel analysis of prosody may be directly used, above all, in a system of automatic speech and speaker recognition, synthesis and linguistic analysis. Some of current works on prosody application in language processing for Polish are introduced.

STRESZCZENIE

Niniejsza praca przedstawia problemy pojawiające się w związku z analizą wzorców intonacyjnych w mowie, ich interpretacją i zastosowaniem na potrzeby Technologii Mowy. Liczne funkcje intonacji są omawiane z punktu widzenia akustyki i fonetyki, językoznawstwa oraz psychologii. Wyniki kompleksowej, wielopoziomowej analizy prozodii mogą być bezpośrednio zastosowane przede wszystkim w systemach automatycznego rozpoznawania mowy i mówców, w syntezie mowy, a także w analizie lingwistycznej. W pracy przedstawiono niektóre z prowadzonych obecnie prac nad zastosowaniem prozodii w przetwarzaniu języka polskiego.

1. Introduction

Among many other factors, the prediction and modeling of prosody plays an essential role in affecting the quality of speech recognition and synthesis. Irrespective of the speech synthesis type (e.g. concatenative or unit selection), prosody is important for several reasons: a) intonation has discourse functions (e.g. signaling given/new information; focus), b) errors in the segmental structure are accepted by the listeners to a greater degree than errors in the suprasegmental structure of the utterance; c) erroneous accent placement or incorrect accent type may significantly change the meaning of the utterance or create the impression of unnaturalness.

In speech recognition systems, suprasegmental features are indispensable as a source of information about the syntactic and semantic structures of utterances, but their extraction is difficult and prone to errors [1]. The verification of the nuclear accent position in a phrase and finding the most essential fragments of an utterance from an informational point of view also makes it possible to reduce the time spent examining the lexicon. Paralinguistic and nonlinguistic aspects of suprasegmentals

play a secondary role; however, they can be useful for the initial decoding of the signal and rapid adaptation of the system to the individual voice characteristics.

Speech recognition, speaker identification/verification and synthesis systems are based on automatic learning. By using a "blind" unsupervised statistical learning method which requires neither laborious experiments nor manual transcription of complex structures at the segmental or/and suprasegmental level, these systems try to omit methodological problems related to insufficient knowledge of interactions between various types of information encoded in a signal. The blind learning method, however, does not provide satisfactory and universal algorithms which would work well and be independent of the choice of language material, the speaker's voice, speaking style, and the acoustic qualities of the surroundings. The formation and extraction of invariants for particular types of information encoded in the speech signal is still controversial. To a certain extent invariant extraction is already possible for well articulated speech (e.g. dictated text), but for spontaneous speech it seems to be an unsolvable problem (eg. [2]).

Automatic learning seems to be an optimal solution for narrowly defined applications of Speech Technology and in fact diminishes the barrier between the potential of automatic speech processing and human speech processing only to a limited extent. However, understanding speech variability requires a comprehensive analysis of prosody: understanding of how a prosodic form indicating a certain function in a language or a task can be converted to a functionally equivalent form in another language or a task [3].

Although there exist a large number of different approaches to prosody modeling, no universal methodology has been worked out so far. Classic approaches to modeling prosody are tone-based models, perceptual models, superpositional models and acoustic stylization models [4, 5]. A more recent approach to the problem is based on the use of a corpus of prosodic data to select the best intonation contour for a given linguistic structure [1]. Some of them have already been applied to the prosody of emotional speech with varied success [6], but the field is still open to research. All existing representation systems for intonation have some drawbacks. For a list and a description of some representation systems, see [7].

An extremely useful overview of systems for a practical approach to the description of the intonation system and a discussion of a relation between forms and functions of intonation has been given by Hirst et al. [3].

The question whether the functional units of intonation can be identified, without all the sources of variation being accounted for, formulated by Wiktor Jassem in 1986 [8], remains one of the major problem areas in the processing of spoken data in general, and of spoken dialogue in particular.

The review and a discussion of the functions of intonation for practical needs of Speech Technology is the subject of the following section. The third section is concerned with prosodic information in speech technology, and the fourth section deals with the specific case of speech recognition.

2. Levels of intonation analysis

There is no consensus as to the priorities of multiple functions of intonation. Some researchers emphasize the grammatical functions, whereas others emphasize that gramatical functions are secondary to emotional functions.

2.1. Acoustic Phonetics

At the acoustic level, intonation carries information about physical features of the variability of fundamental frequency, statistical distributions, short and long disturbances, relation to segmental features on physical level (e.g. voiced/unvoiced segment, micro-prosody).

2.2. Linguistics

2.2.1. Prosody

Common to all intonation modeling systems is the division of utterances into prosodically-marked units or phrases, where prosodic marking may include phenomena such as *audible pause* or *filled pause*, *rhythmic change*, *pitch movement* or *reset*, and *laryngealisation*. Dividing an utterance into such units, speech segmentation is usually the first step taken when carrying out a prosodic analysis. The next step is an analysis of accentuation. Words are made prominent by the accentuation of (usually) their lexically stressed syllable.

Many Western European languages have more than one accent type. It is thus necessary to capture not only on which word an accent is realized but also which kind of accent is used. The most important factors for linguistic motivated features are: speaker's attitude, discourse condition, thematic accent placement [9, 10]. Effects of segmental features and the lexically conditioned length of the tone group should be also analyzed.

2.2.2. Discourse functions

The discourse function of intonation regulates conversational behavior. It provides information about what the speaker is doing in speaking, that is whether he is questioning, advising, encouraging, disapproving, etc. It signals when one has finished speaking and whether another person is expected to speak (regulates turn-taking), a particular type of response, etc. An appropriate pitch movement (fall/rise/level) signals if the phrase is closed, finished, definitive/open, not finished/neutral. The F0 register signals differences in emphasis of finality, nonfinality etc.

2.2.3. Sociolinguistic factors

The function of intonation on the sociolinguistic level provides information about regional varieties, sociocultural background and socioeconomic status, and is especially useful for speaker characterization/recognition and speech recognition when taking into account pronunciation variants, on both segmental and suprasegmental levels.

2.2.4. Psycholinguistic and socio-psychological factors

Certain emotional states which can be controlled by the speaker to some extent, are often correlated with physiological states which in turn have quite mechanical, and thus predictable effects on speech, especially on its prosodic structure. For instance, when a person is in a state of anger, fear, or joy, the sympathetic nervous system is aroused and the speech becomes loud, fast and enunciated with a strong high-frequency energy [11]. When one is bored or sad, the parasympathetic nervous system is aroused which results in a slow, low-pitched speech with little high frequency energy. Furthermore, the fact

that these physiological effects are rather universal means that there are common tendencies concerning the acoustic determinants of basic emotions [11–13].

The research frequently reports conflicting results, due to differences in experimental design and/or interpretation of results.

Several pattern recognition techniques have been used for emotion recognition via acoustic speech features or face features such as Neural Networks, Support Vector Machines (SVMs), k-Nearest Neighbours, Hidden Markov Models (HMM). However creating of a corpus of emotional speech data [14] for statistical methods is very serious problem.

The patterns of complex pitch movement: rise-fall/fall-rise mostly correlate with emotions and emphasis (eg. rise-fall – possible correlation – approval, admiration fall-rise – possible correlation – astonishment, disbelief). The shift of pitch register has been shown to indicate meaning. Ohala, labeled this feature as submissiveness [15]. Median or mean F0 is a good predictor for perceived pitch register. These features probably mostly correlate with attitude (suggestion, confirmation of information without agreement, not decided belief, etc) and excitement.

2.3. Psychological and neuropsychological levels

2.3.1. Psychological personal levels

It has long been known that, quite apart from what is said, a speaker's voice conveys considerable information about the speaker, and that listeners utilize this information in evaluations of speaker's attributes. One such type of information consists of cues to the speaker's personality traits, the most fundamental source of variation between humans. Recent work explores the automatic detection of other types of pragmatic variation in conversation, such as speaker charisma, dominance, point of view, subjectivity, opinion and sentiment. Personality affects these other aspects of linguistic production, and thus personality recognition may be useful for these tasks, in addition to many other potential applications. Personality is typically evaluated along five dimensions known as the 'Big Five': extraversion vs. introversion, emotional stability vs. neuroticism, agreeableness vs. disagreeableness, conscientiousness vs. unconscientious, openness to experience [16].

However, to date, there is little work on the automatic recognition of personality traits based on the speech signal.

2.3.2. Neuropsychological (neurocognitive)

Stress in relation to speech needs to be carefully defined, to distinguish it from the meaning of *stress associated with an accented syllable* in linguistics. The term *stressor* will be used to denote a stimulus which tends to produce a stress response produced by individuals and which may vary to a great degree. Therefore the *speech under stress*, which is less ambiguous, will be used. The taxonomy of stressors and methodological aspects of stress detection has been given by Hansen [17].

Physical stressors include vibration and acceleration, disturbances of the respiratory system, eg. personal equipment that includes clothing and other items worn on the body, which may exert pressure on the vocal apparatus. Physical stressors include a wide range of stimuli, from narcotic drugs to sleep deprivation, and many of these may also have second and third order effects.

Perceptual stressors form a much more limited set, but again may also have third order effects arising from, for example, frustration at the difficulty of communicating via a poor channel (e.g. Lombard effect).

Psychological stressors involve a range of stimuli that could be considered almost unlimited, as the input is interpreted in the light of the individual's beliefs. Workload will be a third-order stressor. The uncontrolled emotional state of the speaker will also be of concern (e.g. personal conflict with a close person).

Stress produced in response to the occurrences in the people's surroundings, perceived by them as unusual and impossible to control, belongs to third order stressors, psychological ones [17, 18]. These kinds of stressor have their effect at the highest level of speech production and cause extreme changes both in segmental and suprasegmental speech structures.

In cases of high stress levels, F0 can reach extreme values (female voices may reach up to 700 Hz). Figure 1 illustrates an utterance of a female marked by an extreme stress level.

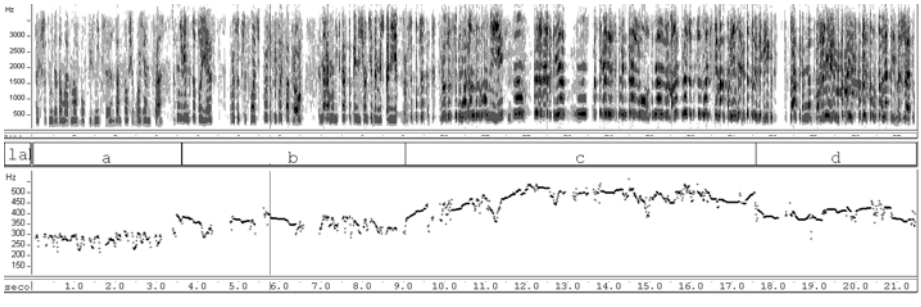

Figure 1: A gradual stress increase in the utterances: a) ***Someone is entering the apartment, (ktoś wchodzi do mieszkania),*** **F_{min} = 220 Hz, b)** ***He's masked, (on jest zamaskowany),*** **F_{min} = 260 Hz, c)** ***he is somewhere [here], (on gdzieś tu jest)*** **– direct threat, F_{min} = 320 Hz, d)** ***Please come to Kwiatowa Street (proszę przyjdź na ulicę Kanałową)*** **– the answer after being asked by a police officer to calm down and tell him the address, F_{min} = 280 Hz.**

At the start the stress level increases even more. It only decreases slightly at the end of the recording after hearing a prompt to calm down. As the stress of the speaker increases, certain processes may be observed: an upward shift in the voice pitch as well as a prominence of the higher frequencies in the spectrum, an increase in the signal's energy and rate changes.

2.4. Multi-level analysis

For the needs of Speech Technology all functions of intonation should be studied at different levels of phonetics, acoustics, linguistics and psychology (Table 1).

The spoken language technologies have achieved outstanding advances in recent years by concentrating on segmental features (speech recognition) and on suprasegmental features only on a linguistic level (synthesis). This is because the current technologies rely mostly on text-reading speech, and only in a very limited way on spontaneous speech. When we try to extend the technologies to more challenging applications (eg. speech

understanding systems) prosodic functions on the highest – neuropsychological – level should necessarily be taken into account.

Table 1: Functions of intonation

Function	Example Categories	Examples
neuropsychological	***neurocognitive***	
Stress indicators	Physical	*Lombard effect*
	Physiological	*Workload*
	Perceptual	*Task-related anxiety*
	Psychological stressors	*Background anxiety*
psychological personalized		
Personality indicators	Extraversion / Introversion	*Expressive speech*
	Emotional stability /	*Arousal*
	Nonstability	*CAPD (Central Auditory*
		Processing Disorders)
Psycholinguistic/ socio-psychological		
Social	Attitudes	*Politeness, irony*
Behavioral	Emotions	*Happiness*
	Relation to content or	*Interests*
	interlocutor	*Disgust*
		Disbelief
Sociolinguistic		
Linguistic	Education	*Formal/colloquial*
Competence	Style	*Foreign influence*
	Dialects	*Environmental influences*
Discourse linguistic		
Discourse	Dialogue acts	*Attraction of attention*
communication		*Turn taking/holding continuation*
Linguistic		
Grammatical	Segmentation (syntactic units)	*Prosodic words, phrases*
	Prominence	*Paragraphs*
	Emphasis (informational units)	*Given/New*
		Focus/Parenthesis
		Theme/Rheme
Acoustic phonetic		
Acoustical speaker	F_0 statistics	*F_0 movements on consonants*
characterization	Microprosody	*jitter laryngalization*
	F_0 disturbances	

3. Prosodic information in Speech Technology

3.1. Speech synthesis

For the purpose of prosody modeling in the Polish module of BOSS (Bonn Open Source Synthesis), which is a corpus based speech synthesis system [19], only fundamental

types of prosodic information, such as lexical stress, pitch accent type and prosodic phrase type, were distinguished.

A pitch accent can be induced by two different mechanisms: a jump to a new pitch level in the syllable nucleus, and a change within the syllable nucleus. The use of a jump rather than a glide or vice versa is often dependent on the makeup of the syllables over which they are spread. If it is only one syllable, a glide is more likely to be used.

- If the pitch accent falls on a syllable with a short vowel, particularly when followed by voiceless consonant, and there is a following syllable, the pitch movement from the accent is more likely to be realized as a jump.
- In general, the use of a jump where a glide might be expected sounds abrupt, whereas the use of the glide when a jump is expected sounds soothing or reproachful. Different distributions and meanings may be found in different languages (English prefers glides, German jumps. Allan Cruttenden [20].

Differences between accent realizations are related to semantic function.

With a view to simplifying the annotation of the pitch accents we took into consideration only two features: direction of the pitch movement and its position with respect to accented syllable boundaries.

The resulting inventory of pitch accent labels includes two labels reflecting pitch movement direction, i.e. falling intonation (HL) and rising intonation (LH). In both cases the movement is realized on the post-accented syllable and the maximum/minimum occurs on the accented syllable. Another three labels also reflect the pitch movement direction (falling, rising and level), but the pitch movement is fully realized on the accented syllable. Level accent is realized by duration. A special label describes the rising-falling intonation on accented syllable (RF).

A perceptual test of the quality of the synthesis showed needs for more precise pitch accent shape modeling [21].

3.2. Computer-assisted language learning (CALL)

Current research in the field of CALL focuses on a more effective integration of computer technology into the learning and teaching languages and on including the prosodic factor into the process of language learning. The main goal is, therefore, to integrate both segmental and suprasegmental aspects, especially in discourse and interaction, and to suggest a complex framework for studying foreign language pronunciation such as the AzAR3.0 [22] software.

Due to its inherent complexity, lack of knowledge about adequate prosody processing both for linguistic and technological needs, as well as lack of attention to the ensuing difficulty in their acquisition, problems of intonation and other prosodic phenomena like rhythm and voice quality were ignored in language teaching for many years.

The main shortcomings of hardware and software used currently for prosody training can be summarized as follows:

- Technical aspects: no extrapolation for voiceless sounds, not entirely correct/reliable F0 extraction, lack of voice quality visualization,

- Methodological aspects: lack of user-friendliness i.e. learners do not know how to interpret displays and evaluate results, lack of integration of such prosodic features as accent, tone, duration, loudness, lack of voice quality analysis – even when a learner can produce individual sound segments, which are very similar to those produced by the teacher, they may still sound 'wrong' due to overall voice quality.

Apart from these points, attention should be paid to acoustic features involved in the realization of intonation. For example, the software could (a) instruct learners to compare the steepness of their falling or rising pitch movement to that of the native speaker, and/or (b) provide a quantitative measurement of the actual pitch slopes of both the native speaker and the learner. An effective feedback of this kind requires implementation of some kind of pitch stylization and normalization. The *Pitch Line* program [22], designed for approximation and parametrization of intonation contours, responds to these needs and could be implemented in the AzAR3.0 environment.

4. Speech recognition

The lexical stress patterns annotated in many different pronunciation dictionaries are effective indicators of pitch accents in speech. This observation should be used to augment the standard ASR (Automatic Speech Recognition) models to improve recognition performance. In particular, it would be of great importance for languages with fixed lexical accent position (e.g. Polish has lexical accent on the penultimate syllable, and Hungarian on the initial syllable).

Polish accent is realized by intonation. There is, as probably in every intonation language, a finite number of melodic patterns, each pattern forming an intonation unit. This unit includes, in Polish, exactly one nuclear tone, which is in the final position. It may be preceded by one or more strong pre-nuclear tones. Nuclear and pre-nuclear tones can be described in terms of the course of F0 and the alignment of F0 variation with syllabic cores. By developing a relatively small number of mathematical formulations related to the F0 course and its alignment with the syllables, it is possible to train artificial neural networks to detect and classify the various tones in the spoken signal and by the same token to detect the pitch accent [21].

The task of integrating prosody within an ASR framework has been dealt with previously [22, 23]. This is a complex problem since prosodic events occur over larger structures, which may lead to ambiguous results.

In our LVCSR (Large Vocabulary Continuous Speech Recognition) system, as first step, the influence of lexical stress information on an ASR system has been investigated by directly modeling selected phonemes placed within stressed syllables.

It was decided to divide each of six Polish vowels /i, y, e, a, o, u/ into two separate monophones, one representing lexically stressed instances and the other representing their unstressed counterpart [23]. The modification was made on the dictionary level only, i.e. no acoustical analysis of stress was performed either on the training set or during the recognition. The applied accentuation rules were based on the main principles for Polish (the stress on the penultimate syllable assumed as default, complemented with a list of accent exceptions). The difference between the standard (39 phonemes) and the

vowel-distinguishing setup is highly statistically significant (matched-pairs difference 3.28%, standard deviation 0.65%).

4.1. Speaker recognition and characterization

As far as automatic speaker verification is concerned, state-of-the-art techniques use a number of parameters drawn from acoustic analysis of the voice such as cepstral coefficients, which must then be employed in a statistical model of voices which is trained from numerous utterances before any verification can be reliably performed. Such methods behave as a black box. But it is necessary to turn our attention to speech structures as these must be dealt with in a more detailed and precise way while being used in general classification tasks. The relation between the speaker's current physical, emotional and cognitive state and the correct dialect, social class markers and speech habits are largely unexplored. For this reason, higher level features that describe a person's expressiveness, as induced by the communicative intention, the current speaker state, the environmental condition, the relationship with his/her interlocutor(s), as well as by the person's identity, as specific for gender, personality characteristics, age, language, and cultural membership [6, 24] are needed. Extraction of the voice characteristics, which contribute to speaker identity, should be preserved under transformations to new tasks or languages [25].

Our study concentrates first on the analysis of intonation on the neuropsychological level. For describing how voice stress is manifested in the acoustic and phonetic structure of the speech signal, a few hundred authentic Polish Police 997 emergency phone calls were selected for acoustic evaluation, the basis for selection being a perceptual assessment [26].

The whole set of recordings was automatically grouped into sessions according to the phone number from which the call was made. Using *Transcriber* annotation software a group of students performed a six-level preliminary annotation on a few hundred recordings with similar length. The annotation included: (1) background acoustics, (2) types of dialog act, (3) suprasegmental description such as speech rate (fast, slow, rising, decreasing), loudness (low voice or whisper, loud voice, decreasing or increasing voice loudness), intonation (rising, falling or sudden break of melody and unusually flat intonation), (4) determination of the context (as the recordings come from a police emergency call database, 3 main contexts were discerned: threat, complaint and depression), and metalanguage description which incorporated (5) the time aspect (past, immediate and potential) and (6) descriptions of emotionally colored phrases in which each was assigned values for three dimensions: potency, valency and arousal, where potency is the level of control that a person has over the situation causing the emotion, valency states whether the emotion is positive or negative and arousal refers to the level of intensity of an emotion.

In highly stressful conditions (e.g. panic) a systematic dynamic shift in pitch over one octave and a significant increase in speech tempo was observed.

The material was divided into four groups: G1: male – stress, G2: male – neutral/mild irritation, G3: female – stress, G4: female – neutral.

The acoustical preparation of recordings consisted in the manual removal of the duty officer's voice from the recordings. For the acoustical analysis with 32 MDVP features [15], in the LDA Linear Discriminant Analysis only 9 have been used: Average

(*F0*), Highest (*Fhi*) and Lowest Fundamental Frequency (*Flo*), Fundamental frequency variation (*vF0*), Jitter (*Jitt*), Amplitude perturbation Quotient (s*APQ*), Degree of subharmonic Segments (*DSH*), Noise to Harmonic Ratio (*NHR*), Degree of voiceless elements (*DUV*). Basic statistical measurements for stressed and neutral speech run over the database showed the relevance of the arousal and potency dimension in stress processing. In states of depression a systematic down shift in pitch and significant decrease in speech tempo was observed.

The LDA analysis of 9 parameters enabled the classification of four groups with an average of 80% accuracy, for two groups (neutral and stressed speech, males and female together) the accuracy was a bit higher, 84%. The results showed that extreme stress can be clearly identified by using only the amplitude information with mean and minimum F0 values.

5. Conclusion and outlook

The basic methodological aspect of this study which requires further consideration suggests that speech processing methods should be creative and based on observation of nature.

Further development of speech technology both in theory and practice requires a comprehensive approach to the problem of signal variability modeling not only on the lowest phonetic-acoustic level, but also on higher, cognitive levels connected with the interpretation of linguistic and paralinguistic information is needed.

References

[1] Hirose K. 2008. Speech Prosody in spoken Language Technologies. *Journal of Signal Processing*, vol. 12, No 1.

[2] Ishi, C.T., H. Ishiguro, N. Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, 531–543.

[3] Hirst, D. and A. di Cristo, eds. 1998. *Intonation Systems: A survey of 20 languages*. Cambridge: Cambridge University Press.

[4] Sagisaka, Y., N. Campbell and N. Higuchi. 1997. *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, New York.

[5] Kohler, K. J. 1997. Modelling prosody in spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi, eds., *Computing Prosody*. New York: Springer, 187–210.

[6] Vidrascu, L. and L. Devillers. 2005. Detection of real-life emotions in call centers. In *Proc. Interspeech* 2005. 1841–1844.

[7] Gibbon, D., I. Mertins and R. Moore, eds. 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. New York: Kluwer.

[8] Jassem W. and G. Demenko. 1986. Extracting linguistic information from F0 traces. In C. Johns-Lewis, ed., *Intonation in discourse*. London: Croom Helm. 1–18.

[9] Hirschberg J. 2002. Communication and Prosody: Functional Aspects of Prosody. *Speech Communication* 36 (1–2), 31–43. http://www1.cs.columbia.edu/~julia/papers/specom00.pdf.

[10] Vaissiere, J. 2004. Perception of intonation. In D. B. Pisoni and R. E. Remez, eds., *Handbook of Speech Perception*. Oxford: Blackwell.

[11] Scherer, K.R. 2005. What are emotions? And how can they be measured? *Social Science Information* 44 (4), 695–729.

[12] Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 169–200.

[13] Cowie, R. and R. R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40, 5–32.

[14] Campbell, N. 2000. Databases of emotional speech, In *Proc. ISCA Workshop on Speech and Emotion*. 34–38.

[15] Ohala, J. J. 1984. An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. *Phonetica* 41 (1), 1–16.

[16] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research* 30, 457–500.

[17] Hansen, J. H. L., C. Swail, A. J. South, R. K. Moore, H. S., E. J. Cupples, T. Anderson, C. R.A. Vloeberghs, I. Trancoso and P. Verlinde 2007. The Impact of Speech Under 'Stress' on Military Speech Technology. NATO Project 4 Report. http://www.gth.die.upm.es/research/documentation/referencias/Hansen_TheImpact.pdf.

[18] Lefter, J., L. Rothkrantz, D. van Leeuwen and P. Wiggers. 2011. Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems* 4 (2), 148–168.

[19] Demenko G., K. Klessa, M. Szymański, S. Breuer, and W. Hess. 2010. Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *International Journal of Speech Technology*. Vol. 13 (2), 85–99.

[20] Cruttenden A., 1986. *Intonation*. Cambridge University press.

[21] Demenko G., ed. 1999. *Analysis of suprasegmentals for speech technology*. Poznań: UAM.

[22] Demenko, G., Wagner, A. and N. Cylwik. 2010. The use of speech technology in foreign language pronunciation training. In *Archives of Acoustics*, 35 (3), 309–329.

[23] Demenko, G., Szymański, M., Cecko, R., Lange, M., Klessa, K. and Owsianny, M. 2011. Development of large vocabulary continuous speech recognition using phonetically structured speech corpus. Proceedings of XVIIth International Congress of Phonetic Sciences, Hong Kong.

[24] Oudeyer, P.-Y. 2003. The production and recognition of emotions in speech: features and algorithms. *Int. J. of Human-Computer Studies* 59 (1–2), 157–183.

[25] Demenko G. and M. Jastrzębska. In press. Analysis of voice stress in call centers conversations. *Proc. of Speech Prosody* 2012.

[26] Shriberg, E., L Ferrer, S Kajarekar, A Venkataraman and A Stolcke. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, Volume: 46 (3–4), 455–472.