Quantification of Segmentation and F_0 Errors and Their Effect on Emotion Recognition

Stefan Steidl, Anton Batliner, Elmar Nöth, and
Joachim Hornegger *

Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung,
Martensstraße 3, D-91058 Erlangen, Germany
stefan.steidl@informatik.uni-erlangen.de

Abstract. Prosodic features modelling pitch, energy, and duration play a major role in speech emotion recognition. Our word level features, especially duration and pitch features, rely on correct word segmentation and F_0 extraction. For the FAU Aibo Emotion Corpus, the automatic segmentation of a forced alignment of the spoken word sequence and the automatically extracted F_0 values have been manually corrected. Frequencies of different types of segmentation and F_0 errors are given and their influence on emotion recognition using different groups of prosodic features is evaluated. The classification results show that the impact of these errors on emotion recognition is small.

1 Introduction

Different types of features have been proposed in speech emotion recognition. In this paper, we focus on prosodic features, which have been proven to effectively discriminate emotional states and are widely used in this field. They model pitch, loudness, and accentuation as well as temporal aspects within suprasegmental units like words or whole utterances. The acoustic correlates are the fundamental frequency F_0, the short-term signal energy, and durations of words, syllables, pauses, etc. A vast number of F_0 extraction algorithms has been developed. For a comparative evaluation see [1]. Nevertheless, all of them are erroneous to some degree. F_0 features are heavily affected by extraction errors; especially octave errors (doubled or halved F_0 values) change the F_0 extrema and the F_0 range significantly. But other features like the slope and the error of the regression line are affected, too. Durations of words or subunits are obtained by a forced alignment of the spoken word sequence to the audio signal. A wrong start and end frame leads to a wrong duration of the word. Furthermore, our data is labelled and classified on word level for which the segmentation is needed as well. In this paper, different types of segmentation and F_0 errors are identified and their

* This work was partially funded by the European Commission (IST programme) in the framework of the PF-STAR project under Grant IST-2001-37599 and the NoE HUMAINE under Grant IST-2002-507422. The responsibility for the content lies with the authors.

published in: Text, Speech and Dialogue (TSD 2008), Proceedings, LNAI 5246, Springer, Berlin, Heidelberg, 2008, pp. 525-534

frequency of occurrence in the FAU Aibo Emotion Corpus, a corpus of spontaneous children's speech in various realistic emotional and emotion-related states, is given. For this reason, the automatic segmentation of the forced alignment and the automatically extracted F_0 values using ESPS have been manually corrected. The impact on emotion recognition is evaluated by comparing the classification performance which results from features calculated with the corrected version with the classification results obtained by features based on the automatic version.

2 The FAU Aibo Emotion Corpus

For this study, the German FAU Aibo Emotion Corpus is used. Here, only a brief description of the corpus is given. More details can be found in [2] and papers quoted therein. The corpus contains speech recordings of 51 children (age 10-13, 21 male, 30 female) of two different schools who were communicating with Sony's pet robot Aibo. The children were led to believe that Aibo was responding to their commands, but the robot was actually being remote controlled by a human operator who caused Aibo to perform a fixed, predetermined sequence of actions. The children were given different tasks like directing Aibo to certain places or through a parcours. To evoke emotions, they were put slightly under time pressure by telling them to direct Aibo as fast as possible through the parcours. At certain predefined situations in the course of the experiment, Aibo did not obey to evoke anger. The task to let Aibo dance was supposed to induce joy. In some tasks, up to three feeding dishes are placed on the carpet. The children were told that one of them contains poison and that they have to make sure that Aibo does not go to this cup under any circumstances. Nevertheless, Aibo approaches exactly this cup in order to elicit slight forms of fear or panic. About 9.2 hours of speech – larger pauses have been removed – have been collected. The recordings of each child have been segmented automatically into smaller 'turns' using a pause threshold of 1 s. Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated each word independently of each other as neutral, which is the default, or belonging to one of ten other classes of emotion-related user states. These categories have been chosen in advance by inspection of the data. Actually, much of the data (48,401 words in total) is neutral. Other states are quite rare (sparse data problem). Hence, a subset of 6,070 words has been selected containing an almost balanced set of the four classes ***A****ngry* (1,557 words), ***M****otherese* (1,223 words), ***E****mphatic* (1,645 words), and ***N****eutral* (1,645 words). The category ***A****ngry* subsumes different but closely related forms of negative attitude like *slight anger*, *touchy/irritated* and *reprimanding*.

3 Manual Correction of the Word Segmentation

A segmentation is necessary to calculate our prosodic features on the word level. If the word boundaries are incorrect, frames outside the word might be consid-

Table 1. Frequencies of different types of segmentation errors

type of error	$\mathfrak{s}1$	$\mathfrak{s}2$	$\mathfrak{s}3$	$\mathfrak{s}4$	$\mathfrak{s}5$	$\mathfrak{s}6$	$\sum$
frequency	300	2,598	1,133	6,249	149	8,427	18,856
	0.6 %	5.4 %	2.3 %	12.9 %	0.3 %	17.4 %	39.0 %

ered for the calculation of the energy and F_0 features while frames inside the word may be missing. Nevertheless, the impact on the energy and F_0 features is supposed to be small since the mean, the extrema, etc. will not change significantly. In contrast, the duration features rely heavily on a high accuracy of the determination of the word, syllable, and phoneme durations which are given by the segmentation. Hence, segmentation errors might have very well an impact on the subsequent emotion recognition. In order to find out how large the influence actually is, the word boundaries have been manually corrected for the complete FAU Aibo Emotion Corpus on the basis of the automatic segmentation obtained by a forced alignment of the spoken word sequence using our own speech recognition system ISADORA [3]. Yet, the exact word boundaries are hard to identify. This is especially true for the end of the word due to reverberation, although a close-talk microphone has been used. The word durations l_{manu} of the manually corrected segmentation and the durations l_{auto} obtained by the forced alignment of the spoken word chain correlate highly (correlation of 0.93). On average, the word in the automatic segmentation is 36.8 frames (frame shift of 10 ms) long – 3.4 frames longer than the average word in the manually corrected segmentation. As this is a systematic error of the aligner which avoids small pauses between words, the impact on the prosodic features is supposed to be small. The two-dimensional histogram in Fig. 1 shows the frequencies of pairs $(l_{\text{manu}}, l_{\text{auto}})$ on a logarithmic gray scale. On average, a word in the forced alignment begins 1.5 frames too early and ends 2.0 frames too late.

In order to have a closer look at the occurring segmentation errors, they are categorised into six groups which are illustrated in Fig. 2. Errors of type $\mathfrak{s}3$ and $\mathfrak{s}6$ indicate that the automatically segmented word is either too short ($\mathfrak{s}3$) or too long ($\mathfrak{s}6$). Automatically segmented words that are shifted slightly to the left or to the right on the time axis, i. e. words that begin and end too early or too late, respectively, but where the automatic and the manual segmentation do overlap to some degree, are of type $\mathfrak{s}2$ ($\mathfrak{s}4$). In the case of no overlap between the automatic and manual segmentation, the words are of type $\mathfrak{s}1$ or $\mathfrak{s}5$ depending on whether the automatically segmented word appears before ($\mathfrak{s}1$) or after ($\mathfrak{s}5$) the manually segmented one. The frequencies of the different error types are given in Table 1. Deviations of at most three frames at both word boundaries are tolerated. Using this threshold, the segmentation of 39.0 % of all words in the corpus is incorrect. In most cases (17.4 %) the automatically segmented words are too long (error type $\mathfrak{s}6$), due to the systematic error of the aligner mentioned above. In about 1 % of the cases, the words are completely misplaced by the automatic alignment (types $\mathfrak{s}1$ and $\mathfrak{s}5$). The average duration of these misplaced words is 264 ms compared to 334 ms of the average word.

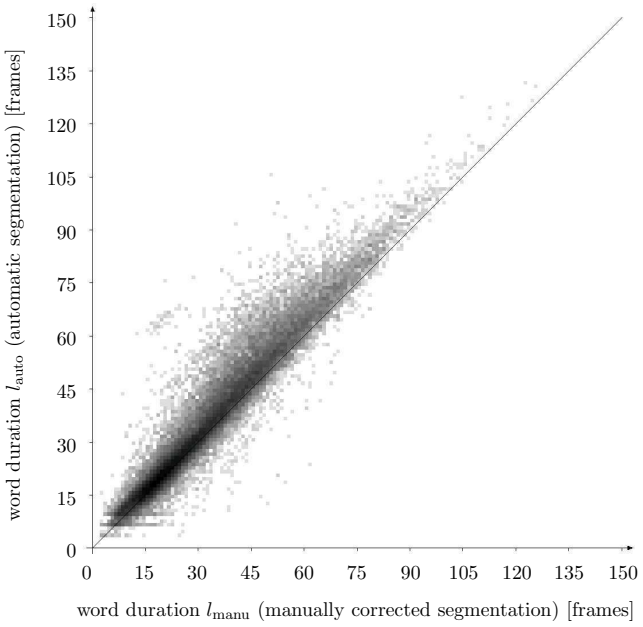

Fig. 1. Comparison of the manually corrected word segmentation with the automatic segmentation of the forced alignment. The histogram frequencies are displayed on a logarithmic gray scale

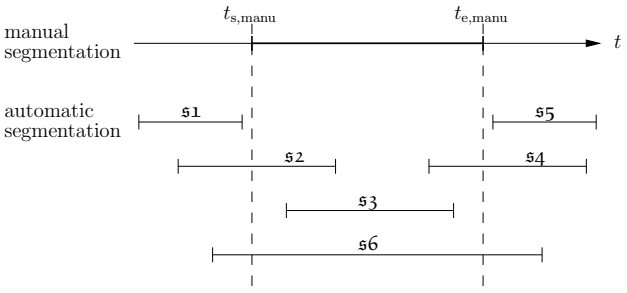

Fig. 2. Different types of segmentation errors

4 Manual F_0 Correction

F_0 features model the rough course of the fundamental frequency. Especially features like the values of the extrema or the range of the F_0 values, but also the regression line and the regression error are directly influenced by F_0 extraction errors. It is an open question to what extent these errors influence the performance of the emotion recognition system. For this reason, the F_0 values of 3,996 turns – the turns that contain amongst others the 6,070 words of the reduced data set described above – have been manually corrected by the second author.

As a reference baseline, the F_0 has been calculated automatically using the freely available and well established F_0 algorithm of the popular *Entropic Signal Processing System* (ESPS) toolkit [4] which is often used for benchmarking. Due to the large amount of data ($> 10^6$ frames), it is impossible to manually determine the length of each period. Hence, the focus is set on the manual correction of obvious errors like voiced/unvoiced errors, octave jumps, or other gross errors. Besides real errors of the pitch extraction algorithm, there are irregularities in the speech production which actually change the fundamental frequency of the signal and can be perceived as suprasegmental irritations modulated onto the pitch contour, but which are not perceived as jumps up or down [5, 6]. Since the manual correction is geared to human perception, a better term instead of 'correction' would be 'smoothed and adjusted to human perception'. We use the term *laryngealisation* for various types of irregular voiced stretches of speech. In [6], five types of laryngealisations have been established: glottalisation, diplophonia, damping, subharmonic, and aperiodicity. The manual correction mostly dealt with the following phenomena:

(1) octave jumps: the ESPS F_0 has been corrected by one octave jump up, in some rare cases also two octave jumps up, or one octave jump down. This concerns rather smooth F_0 curves which had to be transposed. In most cases, it is a matter of irregular phonation where the extraction algorithm modelled pitch rather 'close to the signal' instead of 'close to perception'. In a few cases, however, no clear sign of laryngealisation can be observed. Sometimes, the context and/or the perception had to decide whether an octave jump had to be corrected or not. If the whole word is laryngealised and the impression is low pitch throughout, then laryngealisation is not modulated onto pitch and the F_0 values were kept unchanged.

(2) smoothing at irregularities: the ESPS curve is not smooth but irregular due to laryngealisations or voiceless parts which ESPS wrongly classified as voiced. Here, often the F_0 values between the context to left and the context to the right were interpolated in order to result in a smoothed curve. In case of voiceless parts, the F_0 values were set to zero.

(3) other phenomena like irregularities at transitions which are not necessarily due to irregular phonation: smoothing at transitions is admittedly a bit delicate – when should it be done if the phenomenon is well known, e. g. in the case of higher F_0 values after voiceless consonants. Sometimes, the context and/or the perception had to decide whether an octave jump had to be corrected or not. A typical problem is a hiatus, i. e. the sequence of one word ending in a vowel followed by, e. g., "Aibo". The perception is rather no pitch movement but 'something' modulated onto the pitch curve. In these cases, various F_0 extraction errors can occur: the F_0 values may be set to zero, i. e. the segment is classified as voiceless, octave jumps up or down may occur, the F_0 values may be fully irregular, or values from low to higher may occur. Here, the F_0 was sometimes interpolated, sometimes doubled, or sometimes not corrected (in the case from 'low to higher'). Sometimes, clear criteria for the one or the other solution could not be found, at least not with a reasonable effort. In voiced-unvoiced-voiced se-

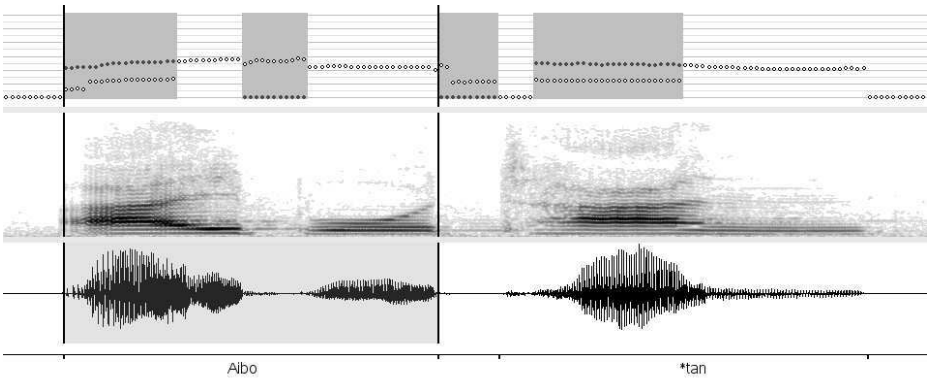

Fig. 3. Manual F_0 correction for the utterance "Aibo, *tanz" (*Aibo, *dance*). The '*' marks word fragments

quences within a word, e. g. in the word "Aibo", the plosive sometimes was set to voiceless even if voiced would have been possible – F_0 postprocessing sometimes interpolates in such cases anyway. In some rare cases, it had to be 'educated guessing' and was not really based on strong criteria.

Fig. 3 shows an example with F_0 correction: below, the time signal, in the middle, the spectrogram, and above, the F_0 values per frame (frame shift of 10 ms). Manually corrected F_0 values are displayed with gray, filled circles. The colour of the background is set to gray if ESPS and manually corrected F_0 values differ. The first part (the `[a]` in `[aI]`) of `/Aibo/` is clearly laryngealised: first glottalisation, then diplophonia, and in the last irregular part, aperiodicity. The intervocalic plosive `[b]` was set to voiceless (note that this is regular in south German dialects). Without using the 'magnifying glass' to scale up the time signal, the `[a]` in `/tanz/` does not display clear signs of irregular phonation.

To illustrate which types of F_0 extraction errors occur how often, a two-dimensional histogram of the pairs ($F_{0,\mathrm{manu}}$, $F_{0,\mathrm{auto}}$) is given in Fig. 4. The frequencies of these pairs are displayed on a logarithmic gray scale in order to make less frequent errors visible as well. Cases where both ESPS and the human corrector decided for voiceless, i. e. pairs (0, 0), are discarded in the histogram due to their very high frequency. The histogram shows that F_0 extraction errors can be categorised into various types of errors. They are denominated with ƒ1 to ƒ7 as defined in Table 2. Since only obvious errors have been corrected, the F_0 values of most frames (94.3 %, s. Table 3) have been kept unchanged resulting in the dark diagonal (ƒ0) in the histogram. Voiced errors (ƒ2), i. e. F_0 values which are considered to be voiceless by the human corrector and voiced by ESPS, result in the vertical straight line. Unvoiced errors (ƒ3), i. e. F_0 values which are wrongly considered to be voiceless by ESPS, yield the horizontal straight line. Three more straight lines result from one (ƒ5) or two (ƒ6) octave jumps down (one half and one fourth of the manually corrected F_0 value, respectively) or

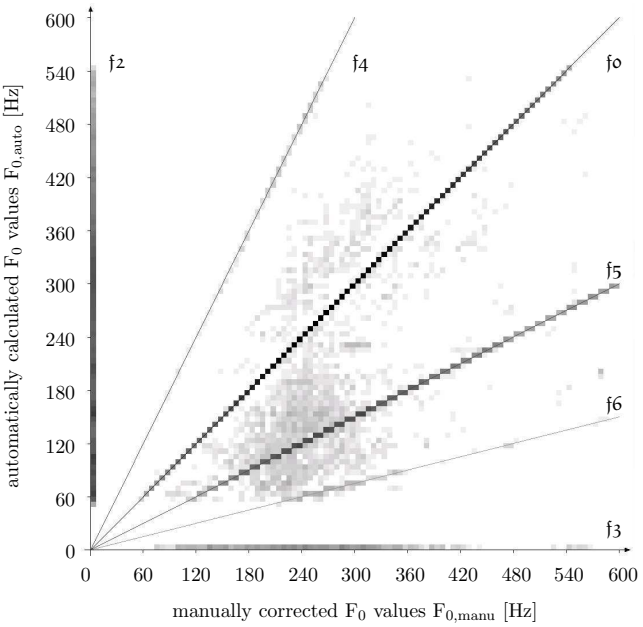

Fig. 4. Comparison of the automatically calculated and the manually corrected F_0 values. The frequencies in the histogram are displayed on a logarithmic gray scale. The straight lines represent identical values (ƒ0) and the error types ƒ2 to ƒ6. Errors of type ƒ1 and ƒ7 are located between these lines

one octave jump up (ƒ(4), ESPS F_0 value is twice the manually corrected one). Other gross F_0 errors are located between these lines.

The frequencies of the different error types are given in Table 3. Numbers are given for the evaluation on the whole turns and for the evaluation only within words. As our prosodic features are word based, F_0 values outside words (45 % of all frames) are irrelevant for our feature extraction. The comparison reveals that – as expected – almost all F_0 errors occur within words. Only voiced errors appear mostly outside words (73 %). The table also lists a few minor errors defined as deviations of less than 10 %. As explained above, minor errors were not in the focus of our manual correction. Anyway, state-of-the-art F_0 features only model the rough course of the fundamental frequency. Thus, minor errors are highly unlikely to influence the emotion recognition.

5 Prosodic Features

We use a set of 95 relevant prosodic features modelling duration, energy and F_0. The latter two groups of features model the course of the energy and the F_0, respectively, within a certain context. Additionally, 30 linguistic features (part-of-speech features) are used. The context can be chosen from two words before

Table 2. Description of different F_0 error types

type	short description	long description
f0	identical	F_0 value calculated by ESPS is not changed by the manual correction
f1	minor error	deviation of the ESPS F_0 value from the manually corrected F_0 value is less than 10 %
f2	voiced error	ESPS calculates a F_0 value for a frame which is considered to be unvoiced by the manual correction
f3	unvoiced error	a frame which is considered to be voiced by the manual correction is marked as unvoiced by ESPS
f4	octave error ↑	ESPS F_0 value is twice the manually corrected F_0 value with a tolerance of 10 %
f5	octave error ↓	ESPS F_0 value is half the manually corrected F_0 value with a tolerance of 10 %
f6	octave error ↓↓	ESPS F_0 value is one fourth of the manually corrected F_0 value with a tolerance of 10 %
f7	other gross error	deviation of the ESPS F_0 value of more than 10 % but not one of the octave jumps mentioned above

Table 3. Frequencies of the different F_0 error types evaluated on the whole turn or only within words

	evaluation			
type of error	whole turn		only within words	
f0 identical	1,050,450	94.3 %	574,485	93.7 %
f1 minor errors	455	0.0 %	452	0.1 %
f2 voiced errors	32,774	2.9 %	8,804	1.4 %
f3 unvoiced errors	1,884	0.2 %	1,877	0.3 %
f4 octave errors ↑	247	0.0 %	239	0.0 %
f5 octave errors ↓	23,718	2.1 %	23,498	3.8 %
f6 octave errors ↓↓	375	0.0 %	364	0.1 %
f7 other gross errors	3,634	0.3 %	3,559	0.6 %

and two words after the actual word. Thus we model, so to speak, a 'prosodic five-gram'. A full account of the prosodic features is beyond the scope of this paper; details are given in [7].

6 Experimental Results

For our 4-class classification problem, cf. section 2, we use artificial neural networks (ANN), implemented within the software package SNNS [8]. A leave-one-speaker-out procedure is employed using 40 speakers for training, 10 speakers for validation and the remaining one for testing in each of the 51 runs. Classification results are obtained for the whole data set and are speaker-independent. In the training and the validation set, the samples of less frequent classes are upsampled to get a balanced set. The features are mapped onto a range from -1 to $+1$

Table 4. Classification results (average recall) for different types of prosodic features

features	manual correction: no	manual correction: yes
all features (PCA: 125 → 95)	61.3 %	61.4 %
F_0 features without position (26)	46.5 %	47.8 %
energy features without position (31)	55.4 %	56.4 %
duration features (17)	52.7 %	52.8 %
duration and position of energy/F_0 (30)	54.6 %	53.8 %
pause features (8)	34.9 %	34.1 %
POS features (30)	52.4 %	

and decorrelated using principal component analysis (PCA). In each run, the topology of the net and two parameters of the training algorithm (weight decay and random seed for initialisation of the network parameters) are optimised on the validation set. The ANNs consist of the input layer containing one node for each feature, one hidden layer of a varying number of nodes, and one output layer of four nodes – one for each class. In Table 4, the classification results are given in terms of the unweighted average recall over all four classes.

Six sets of feature groups are evaluated: The first set contains all prosodic features. In order to reduce the computational costs in the leave-one-speaker-out procedure, the number of features is reduced from 125 to 95 features using PCA. The second and the third set contain 26 F_0 features and 31 energy features only, respectively. Features describing the position of the F_0 and energy extrema are excluded since position features model the duration between the extremum and the reference point. Hence, they are regarded as duration features. The fourth and the fifth set are duration features. In the latter, the position features of the F_0/energy extrema are included. The sixth set contains eight features describing filled and unfilled pauses between words. The last set contains 30 part-of-speech features. For all subsets as well as for the combination of them, the changes caused by segmentation and F_0 errors are not significant. The generally lower relevance of F_0 features, in comparison with energy and duration features, is in line with other studies, cf. [2] and [9].

7 Discussion and Concluding Remarks

Our corpus has been labelled and classified on the word level which is quite unique in research of emotion. This approach takes into account that emotion-related states can change rather quickly – even within utterances. Nevertheless, the opportunity to merge words into larger units like chunks or turns and to map emotion labels from word level onto these larger units is still possible. This approach has been pursued in [2]. One might as well argue that the influence of the described errors on emotion recognition also depends on the choice of features. In [2], results from our initiative CEICES are presented where F_0 and duration features of the participating research institutes are combined covering

a plethora of state-of-the-art prosodic features. Units of analysis were syntactically/semantically meaningful 'chunks' with 2.9 words per chunk on the average. The classification results confirm as well the low impact of F_0 errors on emotion recognition. Note that matters were different for the two-class problem prominence where erroneous F_0 values yielded significantly lower classification performance. However, most important features were different: for corrected values, features modelling the slope (regression) were more important whereas for automatically extracted features, the more robust mean came to the fore. Thus we should not conclude that automatic extraction is generally good enough and does not contribute to classification errors: the automatic segmentation is based on word recognition which is only 'perfect' in forced alignment. In a fully automatic speech recognition system, wrong word recognition can yield wrong segmentation as well, and this in turn might very well contribute to wrong linguistic features for bag-of-words or part-of-speech classes. F_0 errors might not be detrimental only if the feature vector models both specific and more general aspects.

With [2] and the present study it has been shown – to our knowledge, for the first time – that the impact of erroneous automatic extraction of pitch and segmentation on emotion recognition might be negligible, even if their frequency is not (some 4% octave errors within words, and almost 40% incorrect segmentation on the word level). This outcome makes it more likely that the difference in recognition relevance, observed for different feature groups, is not due to some 'surface phenomena' such as the extent of erroneous extraction.

References

1. de Cheveigné, Alain and Kawahara, H.: Comparative Evaluation of F_0 estimation algorithms. In: Proc. Eurospeech 2001, Aalborg, Denmark, pp. 2451-2454
2. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L. and Aharonson, V.: The Impact of F_0 Extraction Errors on the Classification of Prominence and Emotion. Proc. ICPhS 2007, Saarbrücken, Germany, pp. 2201–2204
3. Stemmer G.: Modeling Variability in Speech Recognition. Logos Verlag, Berlin, 2005
4. Talkin D.: A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B., Paliwal, K. K. (eds.): Speech coding and synthesis, Elsevier Science, 1995 (495–518)
5. Batliner, A., Steidl, S. and Nöth, E.: Laryngealizations and Emotions: How Many Babushkas? In: Proc. of International Workshop on Paralinguistic Speech - between Models and Data (ParaLing 2007), DFKI, Saarbrücken, Germany, 2007 (17–22)
6. Batliner, A., Burger, S. and Kießling, A.: MÜSLI: A Classification Scheme For Laryngealizations. In: House, D., Touati, P. (eds.): Proc. of an ESCA Workshop on Prosody, Lund University, Lund, Sweden, 1993 (176–179)
7. Batliner, A., Fischer, K., Huber, R., Spilker, J. and Nöth, E.: How to find trouble in communication. *Speech Communication*, vol. 40, 2003 (117–143)
8. Zell, A., Mache, N., Sommer, T. and Korb, T.: The SNNS Neural Network Simulator. In: Radig B.: Proc. of Mustererkennung 1991, 13. DAGM-Symposium, München, Germany, Informatik-Fachberichte, vol. 290, Springer, 1991 (454–461)
9. Kochanski, G., Grabe, E., Coleman, J. and Rosner, B.: Loudness predicts Prominence; Fundamental Frequency lends little. JASA, vol. 11, 2005 (1038–1054)