Using prosody for the improvement of automatic speech recognition

GYÖRGY SZASZÁK, KLÁRA VICSI

Department for Telecommunication and Media Informatics,
Budapest University for Technology and Economics

{szaszak, vicsi}@tmit.bme.hu

Keywords: speech recognition, prosody, boundary detection, prosodic segmentation

This paper describes sentence, phrase and word boundary detection based on prosodic features, implemented in a HMM-based prosodic segmentation tool. Integrated into a speech recognizer, an N-best rescoring is performed based on the output of the prosodic segmenter, which determines the prosodic structure of the utterance. In an ultrasonography task, we obtained 3,82% speech recognition error reduction using a simplified bi-gram language model.

1. Introduction

Prosody or supra-segmental features are integrant parts of human speech, they provide cues for the listener to understand the meaning by segmenting the speech flow, by emphasizing the important or new information, etc. Moreover, prosody carries sentence mood (modality) and allows the speaker to express emotions, which are embedded acoustically into the speech utterance.

From the point of view of speech technology, high quality speech synthesis would be impossible without modelling prosody, which means definition of the proper intonation, stress and logical segmentation. In speech recognition, however, prosody was not addressed as an information source for a long time, even if supra-segmental features provide not only segmentation information or some representation of nuances in the meaning, but they might by themselves carry information not contained in any other speech related feature. Automatic speech recognizers should exploit this information source in order to ensure some redundancy for speech decoding and also to catch information which would be lost otherwise. For example, automatic classification of sentence modality can be crucial in several speech technology based information retrieval systems, hence several sentences can be composed from identical word chains, the meaning being still different because of the differing sentence mood [1] (question or statement, for example). This is even more important if – like in Hungarian – the subject-predicate inversion does not appear to predict syntactically the sentence modality. In traditional statistical speech recognition, sentence modality classification would be impossible in many cases.

Prosody can, however, be very useful also in traditional speech recognition by providing segmentation information (boundary detection) about the speech utterance. Boundaries of sentences, clauses, syntagms or even some word boundaries can be identified based on supra-segmental features, and the information about the temporal localization of these boundaries can help reduce searching space during the decoding process by removing or penalizing hypotheses not fitting the determined prosodic pattern. Searching space reduction means more robust (more accurate) and faster recognition, recognition speed being one of the critical factor when treating agglutinating languages like Hungarian, Finnish, Turkish, etc. in systems, if real time operation is a basic requirement.

Prosody can also help syntactical and semantic analysis [3] and can predict information-rich segments of speech by detecting stress.

Prosodic features – even if they have not became integral parts of speech recognizers yet – were examined and exploited by several research groups, mainly for English and German languages. Veilleux and Ostendorf elaborated an algorithm rescoring N-best lattices based on prosodic information [10]. N-best lattices are graphs representing recognition hypotheses, each arc having an associated score which functions as a weight, calculated from acoustic and linguistic analysis of the input speech. Based on prosodic information and analysis, these scores can be modified, this is called N-best rescoring. Indeed, it has the same effect as if a prosodic analyser module added his own scores to the acoustic and linguistic ones. The final recognition result is given as the path having the highest score (the most probable path) through the lattice. A similar work was presented for German language in [2].

Gallwitz et al. developed an integrated speech recognizer [1], treating and exploiting "traditional" acoustic and prosodic-acoustic features in parallel. The authors of the present article have also examined the use of prosody in speech recognition [12].

2. Extracting acoustic-prosodic information from speech

For representation of prosody, fundamental frequency (F0), energy level and time course are measurable. Based on our earlier analysis reported in [8], F0 and energy were found to be characteristic when considering emphasis detection. The extraction of prosodic information is performed using the Snack package of KTH [7]. The extraction of F0 is done by AMDF method using a 25 ms long window. The frame rate is set to 10 ms. The obtained F0 contour was firstly filtered with our anti-octave jump tool. This tool eliminates frequency halving and doubling, and also cuts F0 values associated to the first and last frames of each voiced speech segment. This was followed by a smoothing with a 5 point mean filter (5 points cover a window of about 50 ms) and then the log values of F0 were taken, which were linearly interpolated. During the interpolation, two restrictions must be fulfilled. Firstly, interpolation should not affect pauses in F0 longer than 250 ms; secondly, interpolation should be omitted if the initial value of F0 after an F0-gap higher than a threshold value. This threshold value depends on the last measured F0 values and equals the 110% of the average F0 value of the three last voiced frames before the gap (unvoiced period).

These restrictions affecting the interpolation were found necessary because an unvoiced period of length more than 250 ms is likely to be a silence, which should also be detected. On the other hand, interpolation of such a long period would yield only a broad approximation. The reason for the maximal rise criteria of 10% for F0 can be explained in the same manner: firstly a silence (including a breath) is likely, secondly, emphasis is expected to produce also such a rise which should not be smoothed by the interpolation. The threshold values to trigger interpolation were determined empirically. An automatic algorithm for the determination of these values based on speaker specific variables (such as speech or articulation rate, F0 dynamic range, etc.) would also be of interest in the future, but this problem is not issued in the current article.

Energy level values were also extracted using the Snack package, the window size (25 ms) and frame rate (10 ms) were identical to those applied for F0. Energy contour was then filtered by a mean filter. Unlike F0, energy level is a continuous variable, so interpolation is not necessary.

After feature extraction and basic shape conditioning described above, delta and acceleration coefficients are appended to both F0 and intensity streams. These coefficients are computed with a regression-based formula (1). The regression is performed in 3 different steps with increasing regression window length: firstly with a window of ±10 frames, secondly with a window of ±25 frames and finally, a window of ±50 frames is used (*W* in equation (1)). This means that the final feature vector consists of 14 elements (original F0 and intensity data + 3-3 delta + 3-3 acceleration components for both of them).

The formula applied was [9]:

$$d_t = \frac{\sum_{i=1}^{W} i(c_{t+i} - c_{t-i})}{2\sum_{i=1}^{W} i^2}, \qquad (1)$$

where d_t is the delta coefficient at time t c_{t-i} and c_{t-i} are coefficients from the stream to be derivated, *W* is the window length given in the number of frames.

3. Using the prosodic information in the speech recognition process

Prosodic information is used to obtain a broad segmentation of the speech on sentence, clause, syntagm and word boundaries. Feeding this information into the speech recognizer, we except a higher accuracy and the implementation of functions presented in the introduction.

Our algorithm is based on the fact that stress in Hungarian is fixed [2]: if a word is stressed, stress is produced on the first syllable. This makes it possible to handle prosodic information without knowing the underlying word and phoneme sequence. Of course, the final aim is to integrate the processing of phoneme characteristic spectral and prosody affected syntactical information in the speech recognizer.

3.1 Training of an automatic prosodic segmenter

The prosodic segmentation is based on the intonation shape of individual stressed speech segments, separated on word boundaries. As a by-product of this recognition, the temporal location of these boundaries is also available. Boundaries are expected to occur on word boundaries, some of which can also be syntagm and/or clause and/or sentence boundaries at the same time. Please note that intonation now is defined in a more detailed interval than one sentence, as the intonation of the sentence is further split into intonationally coherent segments, so that they coincide with stress and hence by word boundaries. Further in the article, intonation is always regarded as some type of “sentence sub-intonation”.

When determining the set of intonation types, a crucial step is to define classes which are well distinguishable and cover all frequent intonation patterns. To accomplish this, only 6 types of intonation patterns were defined. Silence is the 7th class. The used intonation pattern are listed in *Table 1.*

For training the prosodic segmenter, training samples were selected from Hungarian BABEL speech database [6] (22 speakers, 1600 sentences). This material was segmented based on intonation patterns shown in Table 1. An initial hand-labelling was then extended to

Table 1.
Intonation patterns used for prosodic segmentation

Label	Intonation pattern	Note
me	variable	Sentence onset unit
fe	rise (stress) – falling	Strongly stressed syntactical unit
fs	rise-fall	Stressed unit
mv	falling	Low sentence ending
fv	rising	High sentence or phrase ending
s	floating	Unstressed unit
sil	–	Silence

a computer aided segmentation using a primitive prosodic segmenter trained on hand-labelled data. Hand-labelling was performed relying on F0 and energy contour and subjective impression after listening.

The prosodic segmenter itself is a Markov-model based system whose structure is very close to standard HMM speech recognizers. The 14 dimensional acoustic-prosodic frames are calculated every 10 ms. The number of states is 11 (after optimization) for each intonation pattern class, the linear HMM models were implemented using the HTK package [9].

3.2 Prosodic segmentation process

Automatic prosodic segmentation is carried out using the same algorithms as in speech recognition: first, the acoustic pre-processing is performed, in our case, this is a prosodic-acoustic pre-processing as described in Section 2; then in the decoding stage, Viterbi algorithm is used to obtain the most probable intonation pattern sequence. Hence we use only 7 different pattern classes, and prosodic-acoustic observation vectors are only 14 dimensional and 1 or 2 Gaussians are sufficient for acoustic-prosodic modelling, the decoding process is very fast.

Similarly to a language model in speech recognition, a prosodic grammar is introduced for prosodic segmentation, which specifies the acceptable intonation pattern sequences. This prosodic grammar is relatively severe, but we found empirically that this improves significantly prosodic segmentation performance, while the number of cases where an error occurs due to insufficient generalization capabilities of the prosodic grammar is very low.

The prosodic grammar is given as (using notations from HTK Book [9], p.163):

$$\begin{aligned} &Sentence = \\ &\quad [sil] < [me]\ \{fe \mid fv\ [s]\ \}\ [mv]\ [sil] > sil \end{aligned} \tag{2}$$

Here, '<>' symbol pair refers to one or more, '{}' symbol pair to zero or more repetitions. The '|' symbol denotes alternatives, the '[]' pair encloses optional events. This proto-sequence is interpreted as the prosodic model of a sentence built from intonation patterns.

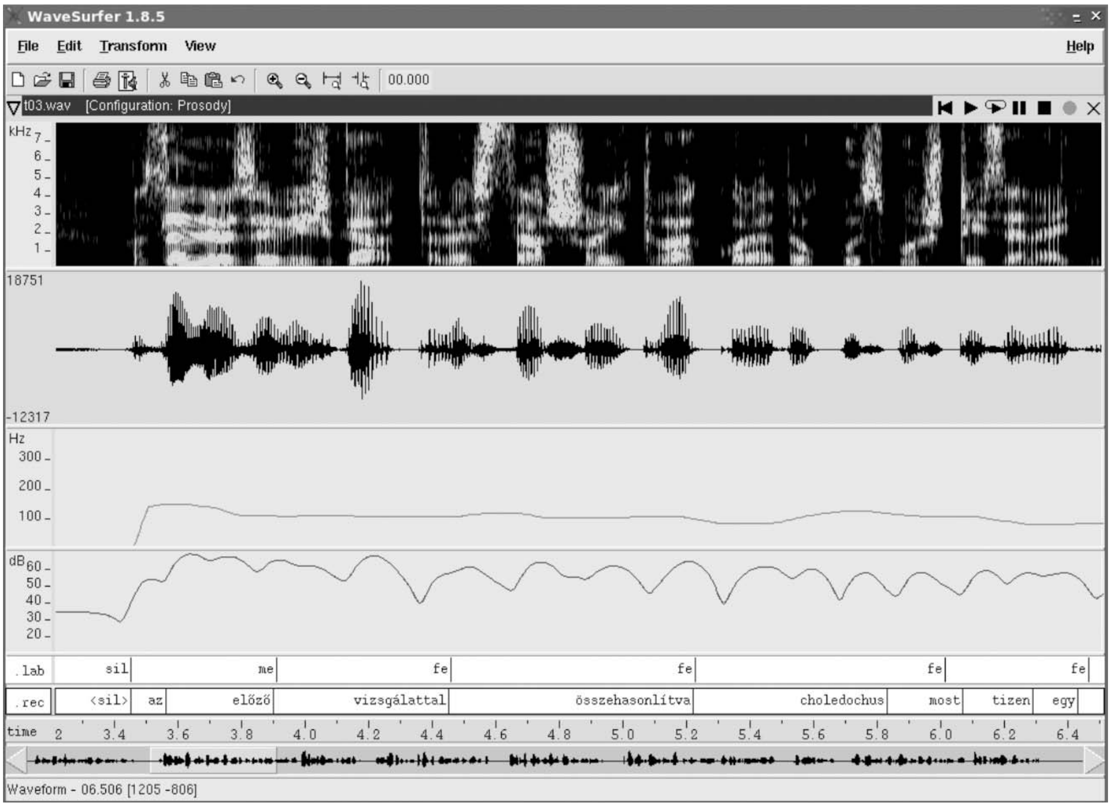
As a by-product of prosodic pattern alignment, the start and ending times of intonation pattern are also calculated. The example shown in *Fig. 1* illustrates the result of the prosodic segmentation process.

3.3 Integration of the prosodic segmenter into the speech recognizer

The output of the prosodic segmenter can be used in speech recognizers to obtain more accurate results and to reduce the searching space. Speech recognizers usually construct a graph (lattice) which specifies the possible outcomes (hypotheses) of the recognition process. Each arc in the graph has its own associated scores (weights) based on a calculation of acoustic and linguistic likelihoods given the input speech signal.

These scores can be re-evaluated (rescoring) with the prosodic information, and so the final recognition result (text output) takes into account prosodic characteristics of the speech. The rescored lattice then goes through the same parsing process as in a standard speech recognizer.

Figure 1. Result of prosodic segmentation for the Hungarian sentence "Az előző vizsgálattal összehasonlítva a choledochus most 11 milliméteres..."



Bounds in the figure from the top to down represent spectrogram (1), waveform (2), interpolated F0 (3), energy (4), prosodic segmentation (5) and underlying word sequence (6).

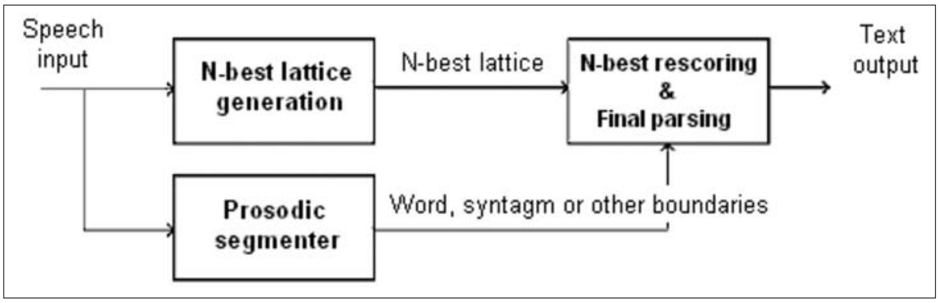

Speech input
N-best lattice generation
N-best lattice
N-best rescoring & Final parsing
Text output
Prosodic segmenter
Word, syntagm or other boundaries

Figure 2. Structure of a speech recognizer with prosodic module

The general recognition process with prosodic module is illustrated in *Fig. 2*.

3.4 Rescoring of N-best lattices

As briefly shown earlier, rescoring of N-best lattices is based on prosodic segmentation. The basic idea is that words or word chains (all recoverable from the N-best lattice) whose syntactical boundaries match well the prosodic structure defined by the prosodic segmentation should be promoted, this means the increase of their associated scores. Similarly, if the temporal characteristics found in the lattice do not fit the prosodic segmentation, the scores can be decreased.

However, the prosodic segmentation might also contain some errors. In spontaneous speech, several characteristic phenomena can lead to even higher prosodic segmentation error rates: mispronounciations, self-corrections, altered prosody or intensive emotions can all disturb the operation of automatic prosodic segmentation. We have already presented a detailed error analysis of the prosodic segmentation in our earlier work [8], now it is sufficient to remember that prosodically predicted boundaries should also be treated carefully when performing lattice rescoring.

As prosodic information is available in the supra-segmental domain, its time resolution is broader than that of the word or phoneme boundaries predicted by the speech recognizer itself. To illustrate this, let's have a look at a final unvoiced fricative of a word: our reference point in prosody is always the last voiced sound (vowel), this uncertainty about F0 is than around the length of a phoneme.

To overcome such difficulties, the locations of syntactic (sentence, phrase, syntagm or word) boundaries (t_B) are transformed to intervals to allow some ΔT time shift when aligning prosodic segmentation to the lattice. Within this interval, the boundary likelihood (L_B) is the highest in the middle and is decreasing towards the limits as defined by:

$$L_B(t) = \begin{cases} A\cos(\frac{\pi}{2\Delta T}t) + C, \; if \; t \in [t_B - \Delta T, t_B + \Delta T] \\ 0 \quad otherwise \end{cases} \tag{3}$$

where A and C are constants. (In our experiments to be presented in Section 4, ΔT was set to 10 frames, which equal 100 ms.) The cosine function was chosen for its simplicity, as it is required that the point to interval transform function has a flat maximum at t_B and decreases towards the limits of the ΔT interval.

The N-best lattice rescoring is then performed as follows. Each edge in the lattice has a word or a word chain associated (with a combined acoustic and linguistic score) and each node has its associated timestamp corresponding to the start and ending times of the word (chain) defined by the edges. A prosodic score is calculated based on the L_B curve, which is the higher if the actual node is the closer (see also Equation 3):

$$Sc_{renum} = w_a L_B(t_{start}) + w_b L_B(t_{end}), \tag{4}$$

where t_{start} is the timestamp of the start node of the word (chain) and t_{end} corresponds to the timestamp of the end node. w_a and w_b are weights.

Hereafter, $L_B(t_i)$ is summed for each frame i of the word (chain) – except the first and last k ones, where t_i is the time index of the actual frame:

$$Sc_{punish} = \sum_{i=k+1}^{N-k-1} L_B(t_i), \tag{5}$$

where N is the total number of frames associated to the word (chain), $k=\Delta T$=100 ms.

The new $Sc_{rescored}$ score of the edge (and so of the word (chain)) is:

$$Sc_{rescored} = w_O Sc_{orig} + w_P(Sc_{renum} - Sc_{punish}), \tag{6}$$

where

Sc_{orig} is the original score, w_O and w_P are weights.

4. Experiment: integrating the prosodic segmenter into an ultrasonography speech recognizer

This section presents an experiment in which the prosodic segmenter functioned as part of a speech recognizer. The integration of the prosodic segmenter into the speech recognizer was carried out as presented in Section 3.3, the operation of the system was the same described in Section 3.4.

The speech recognizer was a Hungarian language, continuous speech recognizer with a 4000 word abdominal ultrasonography dictionary and a corresponding bi-gram language model. This latter was binarized, so it reflected only whether a word sequence was grammatically allowed or forbidden. This reduction was used in order to test the impact that prosodic information can add to speech recognition. However, in large vocabulary speech recognizers such a language model simplification can be useful, as the creation of a language model which covers representatively the application domain is very time and money consuming, mainly for agglutinating languages – like Hungarian –, where even a relatively close application domain needs a larger vocabulary due to the several inflected forms of basic words.

The ultrasonography speech recognizer was implemented in HTK environment, using the “classical” 39 MFC coefficients, 32 Gaussian mixtures for each phoneme state and 10 ms frame rate. For training the 37 acoustic phoneme models, approx. 8 hours of speech was used form MRBA [11] database. The training corpus was segmented on phoneme level.

We integrated the prosodic segmenter into this recognizer in order to analyse recognition performance. The weights in equations (4) and (6) were set as follows: w_a=0,5, w_b=0,5, w_O=1, w_P=2,5.

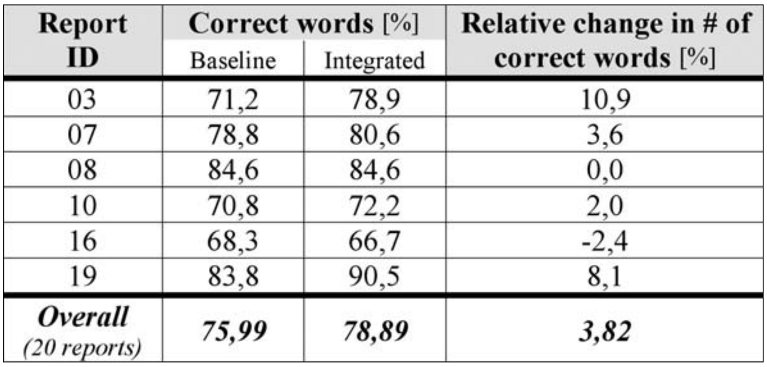
4.1 Results

The testing was carried out on a set of 20 medical reports in the domain of abdominal ultrasonography. (A report contains approx 10 to 20 sentences.) The baseline and the integrated systems worked in an identical environment (same conditions, same recorded reports). Results are presented in *Table 2*. Out of 20, 6 medical reports were representatively selected to be presented in Table 2 in order to allow deeper analysis of results. The overall relative increase in the number of correctly recognized words was 3.82% for the whole test set.

The relative change in the number of correctly recognized words varies from report to report. In case of report ID 03, the relative improvement was over 10%, however, performance might be the same (ID 08) or even worse (ID 16) in the integrated prosodic segmenter system than in the baseline system. Further investigating each medical report and their prosodic segmentation, it was found that a decrease in the performance of the integrated system compared to the baseline one was caused by the errors of the prosodic segmenter, which can be misled by a less proper pronunciation in terms of supra-segmental features, or the error of the pitch detector algorithm can also lead to false boundary detection (prosodic segmentation).

Pitch detectors are sensible to hoarsed (glottalized) speech, some errors were also caused by this phenomenon. On the other hand, reports which were correctly uttered concerning prosody, show a higher improvement compared to the baseline system. A prosodically correct utterance does not require per se professional voicing skills, a common, prosodically well formed pronunciation is sufficient.

Table 2.
Ratio of correctly recognized words with baseline system vs. integrated system



Report ID	Correct words [%]		Relative change in # of correct words [%]
	Baseline	Integrated	
03	71,2	78,9	10,9
07	78,8	80,6	3,6
08	84,6	84,6	0,0
10	70,8	72,2	2,0
16	68,3	66,7	-2,4
19	83,8	90,5	8,1
Overall *(20 reports)*	***75,99***	***78,89***	***3,82***

Please note that in our algorithm, syntactical boundaries missed by the prosodic segmenter do not alter recognition performance. Of course, the more syntactic boundaries the prosodic segmentation reveals, the more performance improvement one can expect. Prosodic segmentation will never locate all of the word boundaries within the speech based solely on supra-segmental features, such a task would exceed even humans’ capabilities.

This is why we used rather the *syntactical boundary* expression through the article instead of *word boundary,* but note also that a syntactical boundary is always a word boundary. We regard as proved that word boundary detection based on prosodic features can improve speech recognition performance.

As a general remark, we think that prosodic segmentation is not always as accurate in the temporal domain and in its resolution capabilities as it would be the ideal one to locate syntactic boundaries. However, this problem can be solved by tracking of the phoneme sequence which would allow a compensation of the prosodic structure in case of necessity. (Of course, tracking in speech recognition is always back-tracking with some delay.) For example, we have mentioned in Section 3.4 that unvoiced phonemes at the end of words can evoke an uncertainness concerning the F0 curve. Such a problem could be more efficiently treated if we knew the underlying phoneme structure or at least if we calculated some confidence of the prosodic segmentation based on phoneme context. We are planning to extend our research in this direction in the future.

5. Summary

Our article addressed the use of supra-segmental (or in other words prosodic) features in speech recognition. We have presented a prosodic segmenter, which aligns syntactical unit assigned intonation patterns or silence to the speech signal. Integrated into an automatic speech recognizer, the prosodic segmenter is used to locate the boundaries of syntactical units, which are also word boundaries. At these boundaries, a prosodic score can be joined by N-best rescoring to the acoustic and linguistic scores available in the speech recognizers. According to our experiments, prosody exploited in this way improves speech recognition performance (and can help the place punctuation marks as well).

We think that the developed prosodic segmenter can also be of interest in natural language processing tools, like syntactic analyzers.

Authors

GYÖRGY SZASZÁK graduated at the Budapest University of Technology and Economics in 2002. In the same year, he became research assistant at the Laboratory of Speech Acoustics, where his main research topics are speech recognition, pronounciation variation, speech database construction and the use of prosody in speech recognition. He is co-author of a dozen of articles and book chapters, mainly in the domain of prosody in speech recognition.

KLÁRA VICSI is the head of the Laboratory of Speech Acoustics at BME-TMIT. She became Doctor of Philosophy in 1992, Doctor of the Engineering Sciences of the Hungarian Academy of Sciences in 2004. She habilitated at BUTE in 2007. She was leader of several Hungarian and international research projects and she is also currently active project leader in the fields of speech acoustics, psychoacoustics, speech recognition and speech databases. She also participates in the development of speech aid systems for hard of hearing children or adults. She has more than 65 Hungarian and international publications and she is author of several book chapters on speech recognition.

References

[1] Gallwitz, F., Niemann, H., Nöth, E., Warnke, V.:
Integrated recognition of
words and prosodic phrase boundaries.
Speech Communication, Vol. 36, 2002,
pp.81–95.

[2] Kassai, Ilona:
Fonetika.
Tankönyvkiadó, Budapest, 1998.

[3] Kompe, R.:
Prosody in Speech Understanding Systems.
LNAI 1307, Springer Verlag, Berlin-Heidelberg, 1997.

[4] Kompe, R., Kiessling, A., Niemann, H., Nöth, H., Schukat-Talamazzini E. G., Zottman, A., Batliner, A.:
Prosodic scoring of word hypothesis graphs.
Proc. of the European Conference on Speech Communication and Technology, Madrid, 1995.
pp.1333–1336.

[5] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A.:
Stochastic pronunciation modelling from hand-labelled phonetic corpora.
In: Modeling Pronunciation Variation for ASR, 1998.
pp.109–116.

[6] Roach, P. S. et al.:
BABEL:
An Eastern European Multi-language database.
International Conf. on Speech and Language, 1996.

[7] Sjölander, K. and Beskow, J.:
Wavesurfer – an open source speech tool.
Proceedings of the 6th International Conference of Spoken Language Processing,
Beijing, China, 2000. Vol. 4,
pp.464–467.

[8] Szaszák, Gy. - Vicsi, K.:
Folyamatos beszéd szószintű szegmentálása szupra-szegmentális jegyek alapján.
In: III. Magyar Számítógépes Nyelvészeti Konf.,
Szeged, 2005.,
pp.360–370.

[9] Young, S. et al.:
The HTK Book (for version 3.3).
Cambridge University, 2005.

[10] Veilleux, N. M., Ostendorf, M.:
Prosody/parse scoring and its aopplication in ATIS.
In: Human Language and Language and Technology Proc. of the ARPA workshop, Plainsboro, 1993.
pp.335–340.

[11] Vicsi K., Kocsor A., Tóth L., Velkei Sz., Szaszák Gy., Teleki Cs., Bánhalmi A., Paczolay D.:
A Magyar Referencia Beszédadatbázis és alkalmazása orvosi diktálórendszerek kifejlesztéséhez.
In: III. Magyar Számítógépes Nyelvészeti Konf.,
Szeged, 2005.,
pp.435–438.

[12] Vicsi, K., Szaszák, Gy.:
Automatic Segmentation of Continuous Speech on Word level Based on Supra-segmental features.
In: International Journal of Speech Technology,
Vol. 8, No.4, 2005.,
pp.363–370.