Automatic Speech Segmentation with the Application of the Czech TTS System

Petr Horák, Betty Hesounová

Institute of Radio Engineering and Electronics,
Academy of Sciences of the Czech Republic, Prague
Chaberská 57, 182 51 Praha 8, Czech Republic
horak@ure.cas.cz, betty@ure.cas.cz

Abstract. This article presents automatic phonetic segmentation of natural speech based on the use of a speech synthesizer and dynamic time warping (DTW) algorithm. The speech synthesizer is used to create a synthetic reference speech pattern with phonetic segmentation information (phonemes, diphones, syllables, intonation units etc.). The reference synthetic speech pattern is then used in the alignment process. The main motivation for this work lay in the lack of usable segmentation tools for Czech, especially for the creation of prosodically labelled databases. The segmentation system has been developed for Czech and it uses the Czech TTS system.

1 Introduction

The automatic phonetic speech segmentation is an important tool for many fields of speech research. It can be used for creation of large prosodically labelled databases for research of natural prosody generation (e. g. by neural nets), for automatic creation of new speech synthesis inventories, and last but not least for the purposes of speech recognition. There are many freely available prosodically labelled natural speech corpora for world languages (i.e. English, French or German), but none for Czech. The new automatic segmentation system should simplify the creation of new corpora. The idea of using speech synthesis for automatic segmentation is not new. Automatic segmentation for French is thoroughly described by F. Malfrère and T. Dutoit in [1]. The developed algorithm is based on the idea of [2] by the same authors sophisticated by G. Strecha in [3] and [4]. Most reports on automatic segmentation are based on using the Hidden Markov Models for TTS alignment. The alignment is basically performed by a trained recognition system. Such recognition systems are typically trained on phoneme models, that is, they acquire some knowledge of the spectral features of phonemes on the basis of many realizations of each phoneme in various phonetic contexts and pronounced by various speakers.

2 Speech synthesis

The Epos speech system [5],[6] is used for creating phonetically labelled synthetic speech. The 8 kHz LPC speech synthesis based on concatenation of 441 Czech and Slovak diphones and vowel bodies is used in Epos for this purpose [7],[8]. The missing information about boundaries between sounds in each diphone was added for this purpose. The Epos with the modified male voice LPC inventory is capable to produce synthetic speech, together with the information about diphones, sounds, syllables and intonation units (similar to words), from written text. The described speech system is illustrated in Fig. 1.

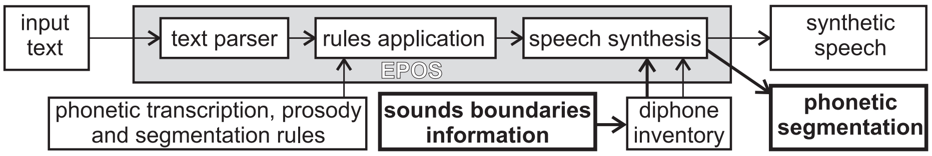

Fig. 1. Enhanced speech synthesis Epos (the bold parts were added).

3 Segmentation

The segmentation algorithm uses separate sentences, therefore a text has to be separated into sentences before the segmentation. The first step of the segmentation consists in the extraction of relevant features from both the reference and the natural speech signals. Both signals are processed frame by frame in each sentence. Six sets of parameters have been used to characterize each speech frame. The first set of parameters defines the representation of the local speech spectral envelope – the cepstral coefficients c_i obtained from linear prediction analysis of the frame [9].

$$c_0 = \ln \sqrt{\alpha}, \tag{1}$$

$$c_n = -a_n - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)c_{n-k}a_k, \quad \text{for } n > 0, \tag{2}$$

where α is linear prediction gain coefficient, $a_0 = 1$ and $a_k = 0$ for $k > M$, M is order of linear prediction analysis. The delta cepstral coefficients Δc_i form the second set of coefficients.

$$\Delta c_0(i) = c_0(i), \tag{3}$$

$$\Delta c_n(i) = c_n(i) - c_n(i-1), \tag{4}$$

where $c_j(i)$ is j^{th} cepstral coefficient of i^{th} frame. The third and fourth sets of parameters are formed by the short time energy and its first difference [10].

$$E(i) = \sum_{m=-\infty}^{\infty} \left(x(m)w(i.N.(1-\mu)-m)\right)^2, \tag{5}$$

$$\Delta E(i) = E(i) - E(i-1), \tag{6}$$

where x is the speech signal, i is number of the frame, N is frame length, m is frame overlapping and M is order of linear prediction analysis, $w(a) = 1$ if $0 \le a < N$, $w(a) = 0$ otherwise.

Finally, the zero-crossing rate and the delta zero-crossing rate coefficients are added to the first four sets of coefficients.

$$Z(i) = \sum_{m=-\infty}^{\infty} f(x(m)x(m-1))w(i.N.(1-\mu)-m), \tag{7}$$

$$\Delta Z(i) = Z(i) - Z(i-1), \tag{8}$$

where x is the speech signal, i is number of the frame, N is frame length, m is frame overlapping, $w(a) = 1$ if $0 \le a < N$, $w(a) = 0$ otherwise, $f(a) = 1$ if $a < k_z(k_z < 0)$, $f(a) = 0$ otherwise.

All of the six parameters sets are normalized into the interval $\langle -1, 1\rangle$. The block diagram of the phonetic segmentation process is shown in Fig. 2.

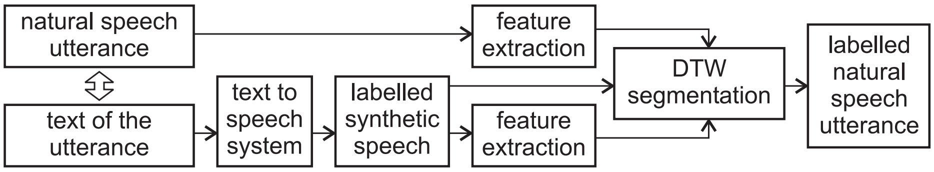

Fig. 2. Phonetic segmentation process.

The second step of the process is the segmentation itself. It is realized with a classical dynamic time warping algorithm with accumulated distance matrix $\mathbf{D}$.

$$\mathbf{D} = \begin{pmatrix} D(1,J) & D(2,J) & \cdots & D(I,J) \\ D & D & \cdots & D \\ \vdots & \vdots & D(i,j) & \vdots \\ D(1,2) & D(2,2) & \cdots & D(I,2) \\ D(1,1) & D(2,1) & \cdots & D(I,1) \end{pmatrix} \tag{9}$$

where I is number of frames of the first signal and J is number of frames of the second signal. This DTW algorithm uses symmetric form of warping function weighting coefficients [11]. The weighting coefficients are described in Fig. 3. In the beginning the marginal elements of the distance matrix are initialized (see equations 10 and 12). Other elements of the distance matrix are computed by the equation 13 [11].

$$D(1,1) = d(x(1), y(1)) \tag{10}$$

$$D(i,1) = D(i-1,1) + d(x(i), y(1)) \quad i = 1 \ldots I \tag{11}$$

$$D(1,j) = D(1,j-1) + d(x(1), y(j)) \quad j = 1 \ldots J \tag{12}$$

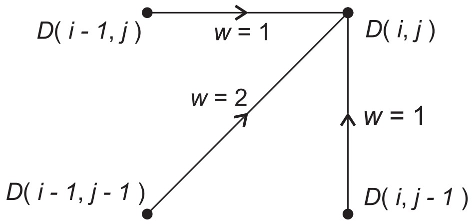

Fig. 3. Weighting coefficients w for symmetric form.

$$D(i,j) = MIN \begin{pmatrix} D(i-1,j) + d\left(x(i), y(j)\right) \\ D(i-1,j-1) + d\left(x(i), y(j)\right) \\ D(i,j-1) + d\left(x(i), y(j)\right) \end{pmatrix} \tag{13}$$

where $d(x(i), y(j))$ is distance between the i^{th} frame of the first signal and the j^{th} of the second signal (see Equation 14), $MIN(x)$ is the minimum function. The distance $d(x,y)$ is a weighted combination of a cepstral distance, an energy distance and a zerocrossing rate distance used to compare a frame from the natural speech signal x and a frame from the synthetic reference signal y.

$$\begin{aligned} d(x,y) = \alpha_0 \sum_{i=0}^{n_{cep}} \left(c_i(x) - c_i(y)\right)^2 + \beta \sum_{i=0}^{n_{cep}} \left(\Delta c_i(x) - \Delta c_i(y)\right)^2 + \\ + \gamma \left(E(x) - E(y)\right)^2 + \delta \left(\Delta E(x) - \Delta E(y)\right)^2 + \\ + \varphi \left(Z(x) - Z(y)\right)^2 + \eta \left(\Delta Z(x) - \Delta Z(y)\right)^2 \end{aligned} \tag{14}$$

An optimization phase over the different parameters used in the distance has led to the following values: frames of 20 ms with a $n = 0.7$ (14 ms) overlap, linear predictive analysis order $M = 8$, $\alpha = 1.5$; $\beta = 1.25$; $\gamma = 1.5$; $\delta = 1$; $\varphi = 1$; $\eta = 1.5$, zerocrossing rate constant $k_z = -20000$.

The next section shows the results of the first experiments performed with our segmentation system.

4 Results

The system presented in the previous section has been preliminary tested and evaluated with one male and one female Czech native speaker. Each speaker pronounced 72 sentences consisting of 3994 phonemes. The automatic segmentation results were compared with manual segmentation of the same data. For the purpose of comparison, Czech sounds were divided into 10 groups according to the manner of articulation and un/voicedness.

The segmentation errors were computed for the beginnings of phonemes and for their lenghts. The errors were divided into 7 categories according to their absolute value in milliseconds. The automatic segmentation results were in most cases better for the female voice although the male speech synthesis voice was used.

The difference between average lengths of manual and automatic labeling was smallest in the group of vowels. Very good results were obtained also in the class of unvoiced explosives, probably owing to their distinct spectrum. Fricatives were, on the other hand, automatically labeled with a marked deviation. The reason might be that initial and final parts of fricatives, as opposed to explosives, are into a greater extent realized on the bordering speech sounds, especially vowels. Considerable differences were found in the group of nasals. This was probably caused by the fact that it is often difficult to distinguish the spectrum of nasals from that of other sonorants.

The automatic segmentation algorithm is very robust. In cases when the natural speech utterance and synthetic speech utterance are not the same, the algorithm skips unequal parts and continues on corresponding parts of speech signals.

5 Applications

The main application of the Czech speech segmentation system is the creation of prosodically labelled speech databases which are needed for further research on prosody modelling, especially for the training of neural nets for the automatic pitch contour generation or for the LPC analysis and synthesis of pitch contours performed in our lab [12].

Another application is the use of automatic segmentation for accelerating the creation of new voices for existing speech synthesizers on the basis of a first set of synthesis units [13]. The phonetic segmentation of speech signals into syllables, sounds (corresponding to phonemes) and intonation units is useful for speech prosody processing while phonetic segmentation to diphones and sounds is helpful for creation of new speech inventories. The research in field of Czech phoneme duration motivated by [14] and [15] was started with the use of the above described automatic speech segmentation.

6 Conclusions

The preliminary evaluation of the automatic segmentation algorithm shows that the accuracy of the automatic segmentation is sufficient for creating prosodically labelled speech corpora and for prosody transplantation, but it is not convenient for a new TTS inventory creation. However, the automatic segmentation algorithm can be used for a new TTS inventory creation supplemented with a manual or semiautomatic adjustment. New speech corpora from several speakers have been recorded. We are working now on the manual labelling of these speech corpora for a better evaluation of the presented system. We plan to use the described automatic segmentation system for creation of a new 16 kHz diphone inventory which could be used for 16 kHz automatic segmentation algorithm. We also plan to extend the new diphone inventory by CC diphones and by two Czech consonants missing in the current 8 kHz diphone inventory (velar "n" and unvoiced "ř").

7 Acknowledgements

This work was supported by the grant No 102/96/K087 Theory and Application of Speech Communication in Czech of the Grant Agency of the Czech Republic and by the Czech Ministry of Education, Youth and Physical Training supply for the COST 258 project.

Special thanks to Guntram Strecha from TU Dresden for his effort on automatic segmentation during his stay in our lab.

References

[1] F. Malfrère, T. Dutoit: Speech Synthesis for Text-To-Speech Alignment and Prosodic Feature Extraction. Proc. ISCAS'97, Hong-Kong, pp. 2637–2640.

[2] F. Malfrère, T. Dutoit: High-Quality Speech Synthesis for Phonetic Speech Segmentation. In: Proc. EuroSpeech'97, Rhodes, Greece, pp. 2631–2634.

[3] G. Strecha: Automatic Segmentation of Speech Signal. Pre-diploma stay final report, IREE Academy of Sciences, Czech Republic, 1999 (in German).

[4] J. Tučková, G. Strecha: Automatic Labelling of Natural Speech by Comparison with Synthetic Speech. In: Proc. the 4th Int. Workshop on Electronics, Control, Measurement and Signals ECMS'99, Liberec, Czech Republic, May 31 – June 1, 1999, pp. 156–159.

[5] J. Hanika, P. Horák: Epos – A New Approach to the Speech Synthesis. In: Proceedings of the First Workshop on Text, Speech, Dialogue–TSD'98, Brno, Czech Republic, September 23–26, 1998, pp. 51–54.

[6] J. Hanika, P. Horák: The Epos Speech System: User Documentation ver. 2.4.42. Available on the web page http://epos.ure.cas.cz/epos.html.

[7] M. Ptáček, R. Vích, E. Víchová: Czech Text-to-Speech Synthesis by Concatenation of Parametric Units. In: Proc. URSI ISSSE'92, Paris, France, pp. 230–232.

[8] R. Vích: Pitch Synchronous Linear Predictive Czech and Slovak Text-to-Speech Synthesis. In: Proc. of the 15th Intern. Congress on Acoustics ICA'95, Trondheim, Norway, 1995, vol. III, pp. 181–184.

[9] J. D. Markel, A. H. Gray jr.: Linear Prediction of Speech. Springer-Verlag, Berlin Heidelberg, New York, 1976.

[10] L. R. Rabiner, R. W. Schafer: Digital Processing of Speech Signals. Bell Laboratories, Incorporated, 1978.

[11] H. Sakoe, S. Chiba: Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In: IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. ASSP-26, 1978, pp. 43–49.

[12] P. Horák: The LPC Analysis and Synthesis of F0 Contour. In: Proceedings of the First Workshop on Text, Speech, Dialogue–TSD'98, Brno, Czech Republic, September 23–26, 1998, pp. 219–222.

[13] T. Portele, K.-H. Stöber, H. Meyer, W. Hess: Generation of Multiple Synthesis Inventories by a Bootstrapping Procedure. In: Proc ICSLP'96, 1996, pp. 2392–2395.

[14] K. Bartkova, C. Sorin: A Model of Segmental Duration for Speech Synthesis in French. In: Speech Communications 6, Elsevier Science Publishers B.V., North-Holland, 1987, pp. 245–260.

[15] Klatt, D. H.: Linguistic Use of Segmental Duration in English: Acoustic and Perceptual Evidence. In: J. Acoust. Soc. Am., Vol. 51, pp. 1270–1278.