33

Automatic Speech Segmentation Based on Alignment with a Text-to-Speech System

Petr Horák
Institute of Radio Engineering and Electronics, AS CR
Chaberská 57, 182 51 Praha 8, Czech Republic
horak@ure.cas.cz

Introduction

Automatic phonetic speech segmentation, or the alignment of a known phonetic transcription to a speech signal, is an important tool for many fields of speech research. It can be used for the creation of prosodically labelled databases for research into natural prosody generation, for the automatic creation of new speech synthesis inventories, and for the generation of training data for speech recognisers. Most systems for automatic segmentation are based on a trained recognition system operating in 'forced alignment' mode, where the known transcription is used to contrain the recognition of the signal. Such recognition systems are typically trained on hidden Markov models of phoneme realisations. Such models are trained from many realisations of each phoneme in various phonetic contexts as spoken by many speakers.

An alternative strategy for automatic segmentation, of use when a recognition system is not available or when there is insufficient data to train one, is to use a text-to-speech system to generate a prototype realisation of the transcription and to align the synthetic signal with the real one. The idea of using speech synthesis for automatic segmentation is not new. Automatic segmentation for French is thoroughly described by Malfrère and Dutoit (1997a). The algorithm developed in this article is based on the idea of Malfrère and Dutoit (1997b) as modified by Strecha (1999) and by Tučková and Strecha (1999). Our aim in pursuing this approach was to generate a new prosodically labelled speech corpus for Czech.

Speech Synthesis

In this study, phonetically labelled synthetic speech was generated with the Epos speech synthesis system (Hanika and Horák, 1998 and 2000). In Epos, synthesis is based on the concatenation of 441 Czech and Slovak diphones and vowel bodies (Ptáček, *et al.*, 1992; Vích, 1995). The sampling frequency is 8 kHz. To aid alignment, each diphone was additionally labelled with the position of the phonetic segment boundary. This meant that the Epos system was able to generate synthetic signals labelled at phones, diphones, syllables, and intonational units from a text. The system is illustrated in Figure 33.1.

Segmentation

The segmentation algorithm operates on individual sentences, therefore both text and recording are first divided into sentence-sized chunks and labelled synthetic versions are generated for each chunk. The first step of the segmentation process is to generate parametric acoustic representations of the signals suitable for aligning equivalent events in the natural and synthetic versions.

The acoustic parameters used to characterize each speech frame fall into five sets. The first set of parameters defines the representation of the local speech spectral envelope – these are the cepstral coefficients c_i obtained from linear prediction analysis of the frame (Markel and Gray, 1976).

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

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

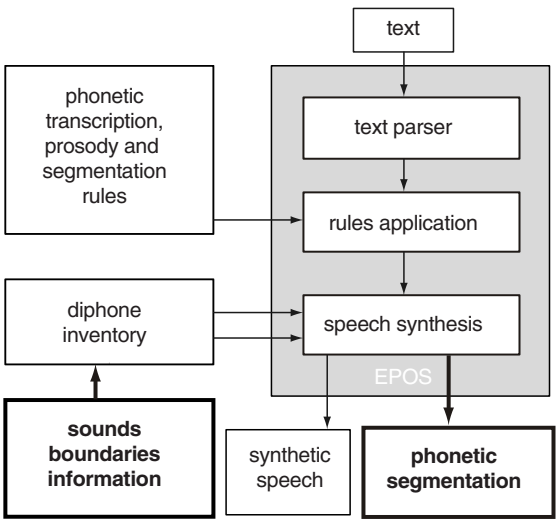

Figure 33.1 Epos speech synthesis system enhanced for segmentation
Note. The bold parts were added

where:

α ... linear prediction gain coefficient
$a_0 = 1$ and $a_k = 0$ for $k > M$
M ... 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 set of parameters is formed by the short time energy and its first difference (Rabiner and Schafer, 1978):

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

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

where:

x ... speech signal
i ... frame number
N ... frame length
μ ... frame overlapping
$w(a) = \begin{cases} 1: & 0 \le a < N \\ 0: & \text{otherwise} \end{cases}$

Finally, the zero-crossing rate and the delta zero-crossing rate coefficients form the last set of parameters.

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

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

where:

x ... speech signal
i ... frame number
N ... frame length
μ ... frame overlapping
$w(a) = \begin{cases} 1: & 0 \le \mathrm{a} < \mathrm{N} \\ 0: & \text{otherwise} \end{cases}$
$f(a) = \begin{cases} 1: & a < k_z \ (k_z < 0) \\ 0: & \text{otherwise.} \end{cases}$

All the parameters are normalized to the interval $\langle 0, 1 \rangle$. The block diagram of the phonetic segmentation process is illustrated in Figure 33.2.

The second step of the process is the segmentation itself. It is realized with a classical dynamic time warping algorithm with accumulated distance matrix **D**.

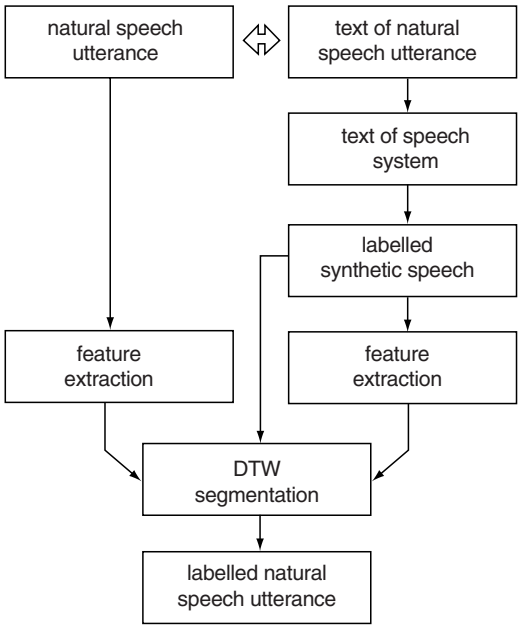

Figure 33.2 Phonetic segmentation process

$$\mathbf{D} = \begin{pmatrix} D(1,J) & D(2,J) & \mathrm{L} & D(I,J) \\ D(1,J-1) & D(2,J-1) & \mathrm{L} & D(I,J-1) \\ \mathrm{M} & \mathrm{M} & D(i,j) & \mathrm{M} \\ D(1,2) & D(2,2) & \mathrm{L} & D(I,2) \\ D(1,1) & D(2,1) & \mathrm{L} & D(I,1) \end{pmatrix} \tag{9}$$

where:

I . . . number of frames of the first signal,
J . . . number of frames of the second signal.

This DTW algorithm uses symmetric form of warping function weighting coefficients (Sakoe and Chiba, 1978). The weighting coefficients are described in Figure 33.3.

In the beginning the marginal elements of the distance matrix are initialized (see equations 10–12). Other elements of the distance matrix are computed by equation 13.

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

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

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

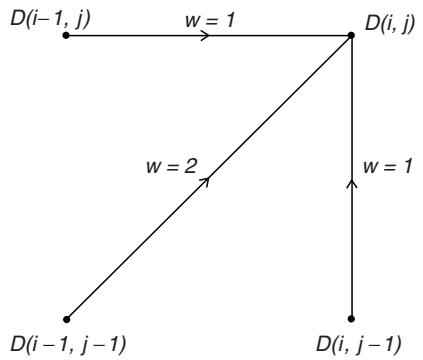

Figure 33.3 Weighting coefficients w for dynamic programming

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

$$i = 1\text{K}I; j = 1\text{K}J$$

where: $d(x(i), y(j))$ ... distance between the i^{th} frame of the first signal and the j^{th} of the second signal (see equation 14) and $MIN(*)$ ... minimum function.

The distance $d(x,y)$ is a weighted combination of a cepstral distance, an energy distance and a zero-crossing rate distance used to compare a frame from the natural speech signal x and a frame from the synthetic reference signal y.

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

Values for the weights in equation (14) and other coefficients of the distance metric were found by an independent optimisation process leading to the following values:

- frames of 20 ms with a $n = 0.7$ (14 ms) overlap;
- linear predictive analysis order: $M = 8$;
- $\alpha = 1.5$; $\beta = 1.25$; $\gamma = 1.5$; $\delta = 1$; $\varphi = 1$; $\eta = 1.5$;
- zero-crossing rate constant $k_z = -20000$.

An example of accumulated distance matrix with minimum distance trajectory is shown in Figure 33.4. The next section shows the results of the first experiments performed with our segmentation system.

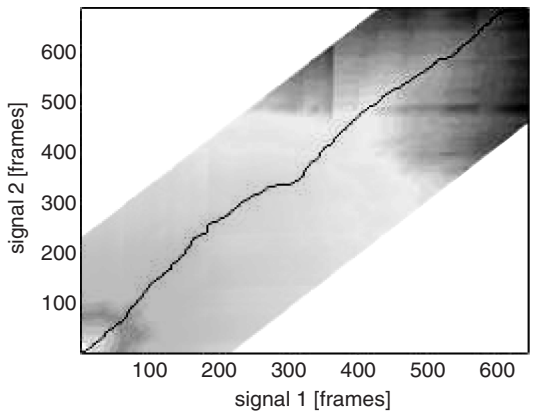

Figure 33.4 An example of a DTW algorithm accumulated distance matrix

Results

The system presented in the previous section was evaluated with one male and one female Czech native speaker. Each speaker pronounced 72 sentences, making a total of 3,994 phonemes per speaker. Automatic segmentation results were then compared with manual segmentation of the same data. Segmentation alignment errors were computed for the beginning of each phoneme and are analysed below under 10 phoneme classes:

vow – short and long vowels [a, ɛ, ɪ, ɔ, o, ʊ, a:, e:, i:, o:, u:]
exv – voiced plosives [b, d, ɟ, g]
exu – unvoiced plosives [p, t, c, k]
frv – voiced fricatives [v, z, ʒ, ɦ, ř]
fru – unvoiced fricatives [f, s, ʃ, x, r̥]
afv – voiced affricates [d͡z, d͡ʒ]
afu – unvoiced affricates [t͡s, t͡ʃ]
liq – liquids [r, l]
app – approximant [j]
nas – nasals [m, n, ŋ, ɲ]

Table 33.3 shows the percentage occurrences of each phoneme class.

Phoneme onset time errors as a function of the absolute value of their magnitude are given in Table 33.1 (male voice). Phoneme duration errors are presented in Table 33.2 (male voice). The same error data for the female voice are given in Tables 33.4 and 33.5. In most cases, segmentation results were superior for the female voice even though the male speech synthesis voice was used (Table 33.6).

As we can see from the tables, the average segmentation error for vowels is the smallest one among all the speech sound groups (see Figure 33.5). Very good results were also obtained for the class of unvoiced plosives. This is probably because the spectral patterning of these sounds is quite distinct, with a clear closure at the onset and with a release that often remains separate from the following speech sound. On the other hand, fricatives showed larger alignment errors. This

Table 33.1 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme onsets for male voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	19.3	37.6	64.5	79.4	86.3	95.2	4.8
n_{vow} (%)	19.1	35.5	61.2	76.0	83.0	95.0	5.0
n_{exv} (%)	21.5	45.0	75.7	87.4	91.2	95.9	4.1
n_{exu} (%)	21.2	42.2	73.5	86.7	93.1	97.9	2.1
n_{frv} (%)	20.9	41.4	68.9	83.6	91.4	95.9	4.1
n_{fru} (%)	11.5	26.2	54.4	74.1	87.2	96.7	3.3
n_{afv} (%)	0.0	10.0	40.0	70.0	70.0	100.0	0.0
n_{afu} (%)	26.1	51.1	75.0	88.0	92.4	100.0	0.0
n_{liq} (%)	18.5	37.6	66.5	85.3	91.2	98.1	1.9
n_{app} (%)	13.5	27.0	52.7	76.4	83.8	92.6	7.4
n_{nas} (%)	22.9	42.5	63.3	73.5	79.3	86.2	13.8

Table 33.2 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme duration for male voice

t[ms]	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	19.4	37.5	64.8	80.0	87.2	90.9	9.1
n_{vow} (%)	19.1	35.7	61.5	76.4	83.7	87.6	12.4
n_{exv} (%)	21.5	45.4	75.7	87.4	91.2	94.0	6.0
n_{exu} (%)	21.2	42.2	73.5	86.9	93.2	95.9	4.1
n_{frv} (%)	21.3	41.8	69.3	84.0	91.8	94.7	5.3
n_{fru} (%)	11.5	26.2	54.8	74.4	87.5	93.8	6.2
n_{afv} (%)	0.0	10.0	40.0	70.0	70.0	80.0	20.0
n_{afu} (%)	26.1	51.1	75.0	88.0	92.4	97.8	2.2
n_{liq} (%)	18.5	37.6	66.8	85.9	92.8	95.0	5.0
n_{app} (%)	14.2	27.7	53.4	77.7	85.8	89.2	10.8
n_{nas} (%)	23.2	43.4	64.9	76.5	83.4	87.6	12.4

may be because the initial and final parts of fricatives, as opposed to plosives, overlap with adjacent speech sounds, especially with vowels. The voiced affricates have the poorest alignment, however there were very few occurrences of these sounds in the corpus. Borders between nasals and other sonorants also showed larger than average alignment error.

The automatic segmentation algorithm seems to be robust to mistakes in transcription. In places where the natural speech utterance and synthetic speech utterance are not the same, the algorithm skips the unequal parts and continues to correctly align the other parts of the signal.

Applications

The main application of the Czech speech segmentation system is the creation of a prosodically labelled speech database to be used for further research on prosody

Table 33.3 Distribution of phoneme occurrences by phoneme class

phoneme class	Number of occurrences	Occurrence (%)
total	4066	100.0
short and long vowels	1736	42.7
voiced plosives	317	7.8
unvoiced plosives	533	13.1
voiced fricatives	244	6.0
unvoiced fricatives	305	7.5
voiced affricates	10	0.2
unvoiced affricates	92	2.3
liquids	319	7.8
approximant	148	3.6
nasals	362	8.9

Table 33.4 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme onsets for female voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	22.4	40.5	66.0	80.3	87.1	95.5	4.5
n_{vow} (%)	20.9	38.6	63.2	77.4	84.1	94.6	5.4
n_{exv} (%)	31.5	55.5	81.1	89.6	94.6	97.2	2.8
n_{exu} (%)	24.0	42.2	70.4	85.4	91.4	97.9	2.1
n_{frv} (%)	29.9	53.7	75.8	86.9	93.0	97.1	2.9
n_{fru} (%)	10.8	21.3	52.1	74.4	87.9	96.4	3.6
n_{afu} (%)	20.0	50.0	80.0	100.0	100.0	100.0	0.0
n_{afu} (%)	32.6	58.7	82.6	90.2	94.6	100.0	0.0
n_{liq} (%)	26.3	48.6	75.2	90.3	95.3	98.7	1.3
n_{app} (%)	23.6	37.8	58.1	76.4	82.4	93.2	6.8
n_{nas} (%)	17.7	30.4	55.2	69.1	76.2	89.2	10.8

modelling, especially for the training of neural nets for automatic pitch contour generation (Horák, *et al*. 1996; Tučková and Horák, 1997) and also for the analysis and synthesis of pitch contours performed in our lab (Horák, 1998). Research into Czech phoneme duration (motivated by Bartkova and Sorin, 1987) was started with the use of the segmentation system.

The speech segmentation tool has also been used for the transplantation of pitch contours between natural and synthetic utterances in order to evaluate our speech-coding algorithm. The block structure of our pitch transplantation tool is given in Figure 33.6.

Future applications for this approach to automatic segmentation will be to accelerate the creation of new voices for existing speech synthesizers on the basis of an existing voice (Portele *et al.*, 1996).

Table 33.5 Error rates (%) as a function of the segmentation magnitude error in ms for phoneme durations for female voice

t(ms)	< 5	< 10	< 20	< 30	< 40	< 50	≥ 50
n_{all} (%)	22.5	40.7	66.4	81.0	88.1	91.9	8.1
n_{vow} (%)	21.0	38.9	63.6	77.9	85.0	89.5	10.5
n_{exv} (%)	31.9	55.8	81.4	90.2	95.3	97.2	2.8
n_{exu} (%)	24.0	42.2	70.4	85.7	92.1	95.5	4.5
n_{frv} (%)	29.9	53.7	75.8	87.3	93.4	95.9	4.1
n_{fru} (%)	11.1	21.6	52.5	74.8	88.5	93.1	6.9
n_{afv} (%)	20.0	50.0	80.0	100.0	100.0	100.0	0.0
n_{afu} (%)	32.6	58.7	82.6	90.2	94.6	96.7	3.3
n_{liq} (%)	26.3	48.9	75.5	90.6	96.2	97.8	2.2
n_{app} (%)	23.6	38.5	59.5	79.1	85.1	89.9	10.1
n_{nas} (%)	17.7	30.4	56.1	71.0	79.3	84.3	15.7

Table 33.6 Average durations of phoneme classes for manual and automatic segmentation for both male and female speakers

Phoneme class	Male speaker		Female speaker	
	manu	auto	manu	auto
total	84.7	88.0	77.0	79.2
short and long vowels	83.3	82.9	81.6	81.3
voiced plosives	77.6	72.9	69.6	65.7
unvoiced plosives	88.4	87.5	78.4	76.9
voiced fricatives	78.2	95.8	66.5	72.7
unvoiced fricatives	113.7	129.8	90.4	105.2
voiced affricates	120.0	141.7	99.9	108.9
unvoiced affricates	126.0	127.3	113.5	115.1
liquids	59.3	63.2	48.4	50.3
approximant	74.0	77.2	61.8	65.6
nasals	87.5	101.1	76.8	88.2

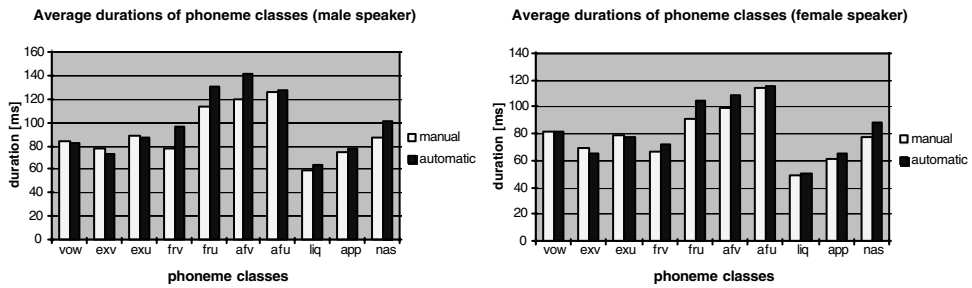

Figure 33.5 Average durations of phoneme classes for manual and automatic segmentation for both male and female speakers

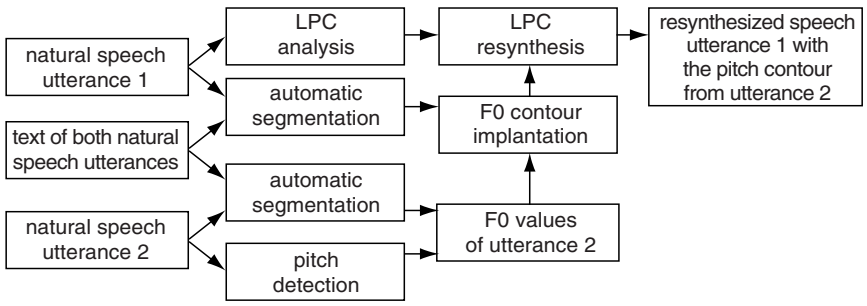

Figure 33.6 Transplantation of a pitch contour using automatic segmentation

Conclusion

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 yet adequate for unit inventory creation. However, the automatic segmentation algorithm could be used for a new unit inventory creation, if 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 Czech consonants missing in the current 8 kHz diphone inventory (ŋ, r̥) (Palková, 1994).

The Epos speech system is a free multilingual speech synthesis system (Horák and Hanika, 1998) which can be used for automatic segmentation of other languages (e.g. German). The Epos speech system can be freely downloaded from http://epos.ure.cas.cz/. We plan to make the automatic segmentation software a free addition to the system.

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 and to Betty Hesounová from our lab for a lot of manual work on segmentation and comparison.

References

Bartkova, K. and Sorin, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication, 6*, 245–260.

Deroo, O., Malfrère, F. and Dutoit, T. (1998). Comparison of two different alignment systems: Speech synthesis vs. hybrid HMM/ANN. *Proceedings of the European Conference on Signal Processing (EUSIPCO '98)* (pp. 1161–1164). Rhodes, Greece.

Hanika, J. and Horák, P. (1998). Epos – A new approach to the speech synthesis. *Proceedings of the First Workshop on Text, Speech and Dialogue – TSD '98* (pp. 51–54). Brno, Czech Republic.

Hanika, J. and Horák, P. (2000). *The Epos Speech System: User Documentation ver. 2.4.43*. Available at *http://epos.ure.cas.cz/epos.html*.

Horák, P. (1998). The LPC analysis and synthesis of F0 contour. *Proceedings of the First Workshop on Text, Speech and Dialogue – TSD '98* (pp. 219–222). Brno, Czech Republic.

Horák, P. and Hanika, J. (1998). Design of a multilingual speech synthesis system. *Sprachkommunikation No. 152, 9. Konferenz Elektronische Sprachsignalverarbeitung* (pp. 127–128). Dresden, Germany.

Horák, P., Tučková, J. and Vích, R. (1996). New prosody modelling system for Czech text-to-speech. In D. Mehnert (ed.), *Studientexte zur Sprachkommunikation. No. 13. – Elektronische Sprachsignalverarbeitung* (pp. 102–107). Berlin.

Malfrère, F. and Dutoit, T. (1997a). Speech synthesis for text-to-speech alignment and prosodic feature extraction. *Proceedings of the ISCAS 97* (pp. 2637–2640). Hong Kong.

Malfrère, F. and Dutoit, T. (1997b). High-quality speech synthesis for phonetic speech segmentation. *Proceedings of the EuroSpeech '97* (pp. 2631–2634). Rhodes, Greece.

Markel, J.D. and Gray, A.H. Jr. (1976). *Linear Prediction of Speech*. Springer-Verlag.

Palková, Z. (1994). *The Phonetics and Phonology of the Czech Language*. Charles University, Prague (in Czech).

Portele, T., Stöber, K.-H., Meyer, H. and Hess, W. (1996). Generation of multiple synthesis inventories by a bootstrapping procedure. *Proceedings of ICSLP '96* (pp. 2392–2395). Philadelphia.

Ptáček, M., Vích, R. and Víchová, E. (1992). Czech text-to-speech synthesis by concatenation of parametric units. *Proceedings of URSI ISSSE '92* (pp. 230–232). Paris.

Rabiner, L.R. and Schafer, R.W. (1978). *Digital Processing of Speech Signals*. Bell Laboratories Inc.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Proc*., Vol. ASSP-26, 43–49.

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

Tučková, J. and Horák, P. (1997). Fundamental frequency control in Czech text-to-speech synthesis. *Third Workshop on ECMS* (pp. 80–83). Université Paul Sabatier, Toulouse, France.

Tučková, J. and Strecha, G. (1999). Automatic labelling of natural speech by comparison with synthetic Speech. *Proceedings of the 4th International Workshop on Electronics, Control, Measurement and Signals ECMS '99* (pp. 156–159). Liberec, Czech Republic.

Vích, R. (1995). Pitch synchronous linear predictive Czech and Slovak text-to-speech synthesis. *Proceedings of the 15th International Congress on Acoustics ICA 95*, Vol. III (pp. 181–184). Trondheim, Norway.