A CONSTRAINED DTW-BASED PROCEDURE FOR SPEECH SEGMENTATION

Peter Vepřek[†] and Michael S. Scordilis[‡]

[†]Electrical and Electronic Engineering Department
The University of Melbourne, Australia

[‡]Wire Communication Laboratory
The University of Patras, Greece

Abstract — Reconfiguring a speech synthesiser to a new voice requires substantial amount of effort. As a result, current synthesisers offer only a very limited number of voices. Methods for automating this process will greatly expand the utility of speech synthesis. This paper presents the development of an enhanced method for the automatic segmentation of speech into phonemes, particularly suited for concatenative speech synthesis. Its effectiveness is tested in an analysis/resynthesis procedure, and in the subsequent perceptual evaluation of typical sentences selected from a large speaker population. Results indicate that this technique can be successfully used for the segmentation of speech for synthesis applications.

INTRODUCTION

Speech synthesis by rule has made considerable advances and it is being used today in numerous text-to-speech (TTS) synthesis systems. Current systems are able to synthesise pleasant-sounding voices at high intelligibility levels. However, because their synthetic speech quality is still inferior to that of fluently produced human speech, efforts are continuing towards the development of natural sounding TTS synthesisers.

An important aspect of speech synthesis is the type of voice modelled during analysis. Although the model speaker is carefully selected, circumstances arise where an inventory of different voices is desirable. Offering an extensive number of different voices will be beneficial for many applications and it will further promote the utility of this important speech technology area. Moreover, in speech-to-speech translation services of the future it will be useful if the synthesiser of the target language could be reconfigured to the voice characteristics of the speaker of the source language. However, currently the development of speech synthesisers is a process that requires substantial resources in time, computation and expertise. As a result, only a small number of voices is available in synthesis systems today.

Speech segmentation is a vital procedure in many speech signal processing applications (Barry and Fourcin, 1992; Vidal and Marzal, 1990; Ljolje et al, 1994). In recent years, efforts have concentrated on developing automatic speech segmentation algorithms mainly as part of speech recognition systems, with varying degrees of success (Torkkola, 1988; Muthusamy and Cole, 1992; Zue et al, 1989; Glass and Zue, 1986 and 1988; Gong and Haton, 1992; Talking 1994).

Automatic speech segmentation is also an integral part of the process of sound inventory selection for synthesis. For the development of speech synthesis the utterances spoken by the model speaker are predetermined and their phonemic transcription is known. Also, the material spoken by the reference speaker has relatively short duration and it is hand-segmented and labelled into its constituent basic units. In order to facilitate the reconfiguration of an existing synthesiser to the acoustic features of a new speaker, and apparent technique is the reading of the same material as used by the model speaker or the reading of a specified passage for subsequent analysis.

Dynamic time warping-based (DTW-based) labelling has been used for the labelling of speech databases for recognition (Gong and Haton, 1992), which has some similarities to the problem investigated in this paper. However, the determination of phonetic deletions, insertions and substitutions is especially important in this application, and the effectiveness of the method must be evaluated both quantitatively as well as qualitatively.

OUTLINE OF APPROACH

Dynamic time warping is based on the derivation of the mapping function between spoken utterances based on a specific distance metric. It is usually followed by the computation of the accumulated distance between a reference prototype and possible candidates, and the subsequent ranking of all candidates in terms of their similarity to the prototype. This technique has been successfully used in many speech recognition applications (Rabiner and Schafer, 1978; Deller et al, 1993; Waibel and Lee, 1990).

The speech material for the signal features evaluated in this paper consisted of the complete SA1 set of the TIMIT database. This set comprised a single sentence, with fixed orthographic transcription, spoken by 630 speakers, of which 438 were male and 192 were female. The sentences were resampled at 12 kHz. Frame-based pre-processing was performed with optimal pre-emphasis, with Hamming windows of 35 ms duration, and a window-shift of 15 ms.

Examination of the phonemic transcriptions of all 630 sentences in TIMIT corpus, and perceptual evaluations were used to determine the single most preferred sentence in terms of intelligibility and phonemic clarity. That sentence was used as the "model" sentence for all subsequent evaluations. The remaining 629 sentences were ranked according to the quantitative degree of similarity to the model sentence. That was achieved by developing an automatic procedure for the comparison of phonemic transcriptions provided by TIMIT. Dissimilarity was derived by counting the number of phonemic deletions, insertions and substitutions occurring in a particular sentence when compared to the model sentence.

QUANTITATIVE EVALUATION

Three sets of scenarios for the segmenter were formulated. In the first set, scenarios A, segmentation performance for different feature vectors was compared. In the second set, scenarios B, two distance measures were evaluated. Finally in the third set, scenarios C, three warping constraints were compared. All scenarios are summarised in Table 1.

Scenario	Feature vector (10 coefficients)	Distance measure	Warping constraint
A_1	PARCOR	Euclidean	none
A_2	cepstrum	Euclidean	none
A_3	mel cepstrum	Euclidean	none
B_1	mel cepstrum	city block	none
B_2	mel cepstrum	Euclidean	none
C_1	mel cepstrum	Euclidean	none
C_2	mel cepstrum	Euclidean	standard duration constraint
C_3	mel cepstrum	Euclidean	phoneme-specific duration constraint

Table 1. Scenarios used to evaluate performance of the segmenter.

Evaluation of the feature vectors

The initial procedure for the development of the segmenter consisted of the implementation and performance testing of different feature vectors. Drawing from work by Davis and Mermelstein (1980), some of the most effective features for recognition were used. They were partial correlation (PARCOR), linear prediction derived cepstrum and mel cepstrum coefficients.

PARCOR coefficients were obtained by linear prediction (LP) analysis using Levinson-Durbin recursion as described in Rabiner and Schafer (1978).

The PARCOR coefficients were then transformed into cepstrum coefficients using (1) resulting in LP-derived cepstrum coefficients.

$$\hat{h}(n) = \alpha_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right) * \hat{h}(k) * \alpha_{n-k}, n = 1 \ldots P, \quad (1)$$

where $\hat{H}(z) = \dfrac{G}{1 - \sum_{k=1}^{P} \alpha_k * z^{-k}}$,

and α_i is the i-th LP coefficient, P is the order of prediction, and G is the gain of the model. $\hat{H}(z)$ and $\hat{h}(n)$ are estimated transfer function and impulse response of the speech production model respectively.

Mel-frequency cepstrum coefficients (MFCC) were calculated using (2).

$$MFCC_i = \sum_{k=1}^{20} X_k * \cos\left[i * \left(k - \frac{1}{2}\right) * \frac{\pi}{20}\right], i = 1 \ldots 10, \quad (2)$$

where X_k is log-energy output of k-th filter. There is a total of 20 mel-frequency spaced filters. The energy spectrum used as input to the filters was obtained by evaluating the transfer function of the LP model of the 20th order along the unit circle.

After the application of the DTW comparison technique, the mel cepstrum coefficients performed the best. The results, in the form of the total number of boundaries for each of the scenarios used, are presented in Table 3. The total number of boundaries for each of the scenarios is classified according to the distance from the labels provided by TIMIT and is shown for the distance ranging from 0 to 9 frames.

Evaluation of the distance metrics

Following the testing of different feature vectors, the city block metric and the Euclidean metric were compared on the same task, for segmentation performance of the mel cepstrum coefficients. The city block metric is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} |x_i - y_i|,$$

and the Euclidean distance metric is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2},$$

where N in the number of features per vector, ten in our case, and **x** and **y** are MFCC vectors.

Evaluation of the two metrics revealed that the Euclidean distance outperformed the city block metric, and as a result, it was adopted as the standard distance measure in all subsequent tests. The results are shown in Table 3.

Evaluation of phoneme duration constraints

The final stage in the development of the system was the evaluation of the effectiveness of duration

constraints. In order to compare this technique against different methods reported elsewhere standard duration constraints (Gong and Haton, 1992) were included in the evaluation procedure.

The standard duration constraints impose penalties proportional to a phoneme duration expansion or contraction whenever that is detected during the comparison of two spoken utterances. That penalty is calculated directly from the warping algorithm by measuring the deviation from the optimum mapping function, which is the diagonal line between the two compared utterances. The penalty was the same for all types of phonemes.

In order to further improve performance, in this work compressibility and expandability was made phoneme-specific. Sentence-initial and sentence-final silences were allowed to vary arbitrarily. Changes to the pause and the epenthetic silence durations were given low and medium penalties respectively. Similarly, changes to the duration of all plosive closures, the glottal stop, the nasals, and the glides were given medium penalties as well. The highest penalty was reserved for the remaining phonemes. Table 2 presents a summary of the phonemic groups and the corresponding penalty levels used in this method.

Group	Phonemes	Penalty
I	h#	no
II	pau	low
III	epi	medium
	bcl/b, dcl/d, gcl/g, pcl/p, tcl/t, kcl/k,	
	q, m, n, w, y	
IV	other	high

Table 2. Groups of phonemes with common compression/expansion penalty.

The different penalty levels resulting from the proposed phoneme-specific cost were incorporated into the accumulated cost as shown in expression (3).

$$AC_{i,j} = \min\left\{AC_{i,j-1}, AC_{i-1,j}, AC_{i-1,j-1}\right\} + LC_{i,j} + \Delta\left(i,j,m,n,phoneme_k\right) \tag{3}$$

where AC is accumulated cost, LC is local cost, and Δ is phoneme-specific duration cost.

The phoneme-specific duration cost Δ was calculated accordingly to (4).

$$\Delta\left(i,j,m,n,phoneme_k\right) = \lambda\left(phoneme_k\right)\cdot\left|\left(m+\left(j-n\right)\frac{L_T-m}{L_R-n}\right)-i\right| \tag{4}$$

where $\lambda \geq 0$. Figure 1 clarifies all variables introduced in (4). The coefficient λ controls how much deviations from *expected* duration are being penalised. Note that when $\lambda = 0$ the accumulated cost collapses into the standard DTW cost.

The local path constraints of the warping process were simple left-to-right, bottom-to-top and left-bottom-to-right-top transitions (Deller et al, 1993).

Evaluation of the three different duration constraints indicated that the phoneme-specific constraints outperform the standard duration constraint. Table 3 summarises the results of this scenario. The system performance was best using the phoneme-specific duration constraint.

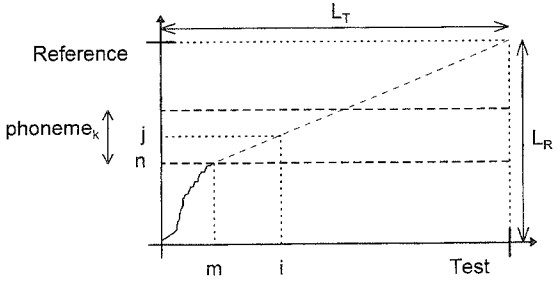

Figure 1. Situation for calculation of the phoneme-specific duration cost.

Scenario	Distance from the TIMIT label [frame]									
	0	1	2	3	4	5	6	7	8	9
A_1	31	39	15	7	4	1	1	1	0	0
A_2	31	40	16	6	3	2	1	0	0	0
A_3	35	39	15	6	3	1	0	0	1	0
B_1	34	39	15	6	3	1	1	0	1	0
B_2	35	39	15	6	3	1	0	0	1	0
C_1	35	39	15	6	3	1	0	0	1	0
C_2	25	31	14	8	5	3	3	2	1	2
C_3	36	40	14	6	2	1	0	0	0	1

Table 3. The total number of boundaries [%] for each of the scenarios.

SPEECH ANALYSIS/RESYNTHESIS AND QUALITATIVE EVALUATION

The goal of the segmentation procedure outlined above was to facilitate the rapid reconfiguration of a speech synthesiser to new voices. Therefore, it was important to specify a method which would make possible to measure the effectiveness of this technique in speech analysis and resynthesis, and to evaluate the perceptual quality of the resynthesised sentences. For this purpose, a concatenative speech synthesiser was employed, based on a 15^{th} order linear prediction model.

For sentence resynthesis the phoneme durational information and the derived fundamental pitch values were provided. The resynthesis strategy employed was based on the representation of the signal by a minimal number of frames, which were as few as one frame per phoneme depending on whether the phoneme was by its nature stationary (e.g. a vowel or nasal) or nonstationary (e.g. a diphthong or plosive). All frames retained represented quasisteady states of the signal. Frame position within the phoneme was first located to be in the middle of the phoneme. The exact frame location was subsequently determined by utilising the first derivative of the short-term energy and of the first and second PARCOR coefficients. The new frame location was placed at the point where at least one of the three features used reached a local minimum. Each frame consisted of the optimal pre-emphasis coefficient used during analysis, 15 PARCOR coefficients and windowed and quantised residual signal. Speech was synthesised from this limited set of frame parameters by interpolating between adjacent frames and switching between appropriate excitation functions at required instances, as provided by the segmenter.

In order to minimise the contribution of secondary factors, phoneme durational and pitch information of the model sentence was used in the resynthesis of all other sentences. Perceptual evaluation of this resynthesis scheme was performed by selecting from the top of the list the five male-spoken and the five female-spoken sentences most similar to the model sentence and from the bottom of the list the five male and five female most dissimilar sentences. The differences of the accumulated boundary errors between these two sets of sentences were derived as shown in Table 4, and it was revealed that the most similar sentence set had been segmented more accurately than the most dissimilar sentence set. However, the error differences are relatively small.

These twenty sentence-pairs were also played in random order. Subjects were asked to listen to each of the twenty natural sentences and to their corresponding resynthesised versions. The speech quality was evaluated on a scale from 0, indicating highest similarity to 10, indicating lowest similarity to the original. After normalisation the average score for similar sentences was 4.85 and for dissimilar 5.37. Results show that although the score difference between the two sets was only 5.2%, the sentences phonetically closer to the model sentence were segmented more accurately than the phonetically dissimilar sentences. Importantly, however, the perceptual effects were not substantial.

Phoneme boundary location error [frame]									
0	1	2	3	4	5	6	7	8	9
1.5%	4.8%	5.1%	1.4%	1.9%	1.4%	0.9%	0.0%	0.0%	0.0%

Table 4. Difference of the accumulated percentage of the total number of boundaries between the sets of most similar and most dissimilar sentences.

CONCLUSION

In this paper, it was shown that phoneme-specific duration constraints improved the performance of dynamic time warping-based speech segmentation. The best segmentation was achieved using 10 mel-frequency cepstrum coefficients with Euclidean distance, together with the phoneme-specific duration constraints. Using this setup 76.0% of the phoneme boundaries were located within +/-1 frame, 90.3% within +/-2 frames, and 98.7% within +/-5 frames. When a simple (non-phoneme-specific) duration constraint is used error rates may increase especially when warped sentences have a variable length leading and trailing non-speech signal included, such as when no proper end-point detection is used. Perceptual evaluation revealed that the perceived difference between the similar set of sentences and the dissimilar set of sentences was only 5.2% which is not substantial. This was in accordance with the quantitative evaluation of the two sets.

REFERENCES

Barry, W. J., Fourcin, A. J. (1992) *Levels of labeling*, Computer Speech and Language, vol. 6, pp. 1-14.

Davis, S. B., Mermelstein, P. (1980) *Comparison of parametric representations for monosyllabic word recognition in continuously spoken utterances*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, pp. 357-366.

Deller, J. R., Proakis, J. G., Hansen J. H. L. (1993) *Discrete-time processing of speech signals*, (New York: Macmillan).

Glass, J. R., Zue, V. W. (1986) *Signal representation for acoustic segmentation*, Proceedings of the First Australian International Conference Speech Science and Technology, pp. 124-129.

Glass, J. R., Zue, V. W. (1988) *Multi-level acoustic segmentation of continuous speech*, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 429-432.

Gong, Y., Haton, J.-P. (1992) *DTW-based phonetic labeling using explicit phoneme duration constraints*, Proceedings of the International Conference on Spoken Language Processing, pp. 863-866.

Ljolje, A., Hirschberg, J. van Santen, J. P. H. (1992) *Automatic speech segmentation for concatenative inventory selection*, Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pp. 93-96.

Muthusamy, Y. K., Cole, R. A. (1992) *Automatic segmentation and identification of ten languages using telephone speech*, Proceedings of the International Conference on Spoken Language Processing, pp. 1007-1010.

Rabiner, L. R., Schafer, R. W. (1978) *Digital processing of speech signals*, (Englewood Cliffs, N.J.: Prentice Hall).

Talking, D. (1992) *The aligner: text to speech alignment using Markov models and a pronunciation dictionary*, Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pp. 89-92.

Torkkola, K. (1988) *Automatic alignment of speech with phonetic transcription in real time*, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 611-614.

Vidal, E., Marzal, A. (1990) *A review and new approaches for automatic segmentation of speech signals*, in L. Torres, E. Masgrau, M. A. Lagunas, eds., Signal processing V: theories and applications, (Elsevier Science).

Waibel, A., Lee, K.-F. (1990) *Readings in speech recognition*, (San Mateo, CA: Morgan Kaufmann).

Zue, V., Glass, J., Phillips, M., Seneff, S. (1989) *Acoustic segmentation and phonetic classification in the SUMMIT system*, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 389-392.