Improving the Accuracy of the Speech Synthesis Based Phonetic Alignment Using Multiple Acoustic Features

Sérgio Paulo and Luís C. Oliveira

L^2F Spoken Language Systems Lab.
INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{spaulo,lco}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract. The phonetic alignment of the spoken utterances for speech research are commonly performed by HMM-based speech recognizers, in forced alignment mode, but the training of the phonetic segment models requires considerable amounts of annotated data. When no such material is available, a possible solution is to synthesize the same phonetic sequence and align the resulting speech signal with the spoken utterances. However, without a careful choice of acoustic features used in this procedure, it can perform poorly when applied to continuous speech utterances. In this paper we propose a new method to select the best features to use in the alignment procedure for each pair of phonetic segment classes. The results show that this selection considerably reduces the segment boundary location errors.

1 Introduction

Phonetic alignment plays an important role in speech research. It is needed in a wide range of applications, from the creation of prosodically labelled databases, for research into natural prosody generation, to the creation of training data for speech recognizers. Furthermore, the development of many corpus-based speech synthesizers [1,2]) requires large amounts of annotated data.

Manual phonetic alignment of speech signals is an arduous and very time consuming task. Thus, the size of the speech databases that can be labelled this way are obviously very constrained, and the creation of large speech inventories requires some sort of automatic method to perform the phonetic alignment. While building a system to automatically align a set of utterances, two different problems can be found. First, we have to know the sequence of phonetic segments observed in those utterances. Then, we have to locate the segment boundaries. The sequence of segments can be obtained by using a pronunciation dictionary or by applying a set of pronunciation rules to the orthographic transcription of the utterances. However, it is, usually, not possible to predict the exact sequence uttered by the speaker and we must take into account possible disfluencies,

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 31–39, 2003.

elisions, allophonic variations, etc. In this work, we will assume that we already have the right sequence of segments and we will focus on the task of locating the segment boundaries.

Several approaches have been taken to try to solve this problem. The most widely explored technique is the use of HMM-based speech recognizers (sometimes hybrid systems, based on HMM and Artificial Neural Networks) in forced alignment mode. This approach relies on the use of phone models built under the HMM framework. This models

are trained using large amounts of labelled data, recorded from several speakers, to take into account the phone's acoustic properties in very different contexts. For single speaker databases, the performance of the system can be improved by adapting the speaker independent models to the speaker's voice. The difficulty of this approach is that it requires the availability of segmented data for the speaker. This material must be annotated following strict segmentation rules so that the resulting system can locate segment boundaries with the necessary precision. When no such system is available, a Dynamic Time Warping (DTW, [3]) based approach can be taken. This technique was used in early days of speech recognition to compare and align a spoken utterance with pre-recorded models, taking into account possible variations on the speaker's rhythm. The recognized utterance corresponded to the model with the minimum accumulated distance after the alignment. The same methodology can be used for the phonetic alignment problem as described in [6] and [7]. This procedure, also known as speech synthesis based phonetic alignment, starts by producing a synthetic speech signal with the desired phonetic sequence that allows us to know the exact location of the phonetic segment boundaries. This can easily be achieved using a modified speech synthesizer. The next step is to compute, every few milliseconds, vectors of acoustic features for both the synthetic and natural speech signals. By using some type of distance measure, the acoustic feature vectors can be aligned with each other using the DTW algorithm. The algorithm result is a time alignment path between the synthetic and natural signal time scales, that allows us to map the segment boundaries from the synthetic signal into the natural utterance. This approach does not require any previously segmented speech from the same speaker but the results depend, in some extent, on the similarity between the synthesizer's and speaker's voice, and they should have, at least, the same gender. The performance of this method is strongly dependent on the selection of the acoustic features used in the alignment procedure and on the distance used to compare them.

This work is part of an effort to automate the process of multi-level annotation of speech signals. A complete view about this problem can be found in [4]. In this paper, we will describe our work on the use of different features to improve the performance of a DTW-based phonetic alignment algorithm. The results of this study lead us to a new method to perform the alignment that uses multiple acoustic features depending on the class of segments to be aligned.

The paper is divided into five sections. The next section describes the process for producing the synthetic reference signal with segmentation marks. The

following section describes an automatic procedure for the selection of the most relevant acoustic features. These results are then applied in the next section, where the alignment procedure is described. The final section compares the results of the new method with a traditional approach.

2 Waveform Generator

An important issue on the DTW-based phonetic alignment is the generation of the reference speech signal. This can be achieved by using some sort of a speech synthesizer, that can be modified to produce the desired phonetic sequence together with the segment boundaries. The problem with this solution is that the signal processing required to impose the rhythm and intonation determined by the prosody module also introduces distortions on the synthetic signal. For our purposes, these prosodic modifications are not necessary and a simple waveform concatenation system was used. Since our goal was to locate the segment boundaries, we used diphones as concatenation units. This way, the concatenation distortion is located in the middle of the phone and does not affect the signal in the phone boundary.

In order to have a general purpose system it must be able to produce any phonetic sequence and the inventory must contain all the possible diphones in the language. We followed the common approach of generating a set of nonsense words (logathomes), containing all the required diphones in a context that minimizes the co-articulation with the surrounding phones. A speaker was asked to read the logathomes in a sound proof room and was recorded using a head mounted microphone in order to keep the recording conditions reasonably constant among sessions. We also asked the speaker to keep a constant intonation and rhythm. The recorded material was then manually annotated.

We used the unit selection module of the Festival Speech Synthesis System[8] to perform the concatenation. A local search is made around the diphone boundaries to find the best concatenation point. We used the Euclidean distance between the Line Spectral Frequencies (LSF) for costing the spectral discontinuities of the speech units.

3 Acoustic Features

We considered some of the most relevant acoustic features used in speech processing: the mel frequency cepstrum coeficients (MFCC) and their differences (deltas), the four lowest resonances of the vocal tract (formants), the line spectral frequencies (LSF), the energy and its delta and the zero crossing rate of the speech signal. Both the energy and the MFCC coefficients, as well as their deltas, were computed using software from the Edinburgh Speech Tools Library [9]. The formants were computed using the *formant* program of the Entropic Speech Tools [10] and the remaining features were computed using our own programs.

Our first experiments showed that each of these features used separately produced uneven results. Depending on the class of phones to be aligned some features proved better than others. For instance, in a vowel-plosive transition, the energy feature was the performer, but for vowel-vowel transition, the best results were achieved using formants as features. This immediately suggested the use of multiple features to distinguish the different phone transition classes.

3.1 Feature Normalization

The combination of multiple features requires a previous normalization step to equalize its influence on the overall alignment cost. It was decided to normalize the values to the range $[0, 1]$.

The first stage was to determine which of the features had values that followed a *Gaussian* distribution. Observing the histograms of each coefficient, the MFCCs and their deltas were the only ones that matched that distribution. The mean and standard deviation were computed for each one of them, and the normalization was then performed, using the equation:

$$x_i = \frac{1}{2} + \frac{X_i - \mu_i}{2\sigma_i} \tag{1}$$

where x_i, X_i, μ_i and σ_i are the normalized value, the non-normalized value, the mean value, and the standard deviation of the i^{th} MFCC, respectively. The LSF values were divided by π. Since the zero crossing rate was computed by evaluating the ratio between the number of times the speech signal crosses the zero magnitude and the number of signal samples existing in a fixed size window (some milliseconds), its values have already the right magnitude (between 0 and 1). For the energy, its delta and for the formants, maximum and minimum values were found for each utterance, and their mean values were computed ($\overline{Y_{imax}}$ and $\overline{Y_{imin}}$). The normalized values were calculated using the following equation:

$$y_i = \frac{Y_i - \overline{Y_{imin}}}{\overline{Y_{imax}} - \overline{Y_{imin}}} \tag{2}$$

3.2 Feature Selection Procedure

Having all the features normalized, the next goal was to find which were more relevant in a given phonetic context. That is, which feature allowed us to locate the boundary with greater precision. For this purpose we had a set of 300 manually aligned utterances that we use to evaluate the relevance of each feature. These utterances were spoken by a different speaker than the one used to record the diphone inventory. The waveform generator previously described was used to produce reference synthetic signals for the phonetic sequences of these utterances and sets of feature vectors were computed every 5 milliseconds for both the reference and spoken signals.

Using the Euclidean distance, a matrix was computed with the distances between all the feature vectors of the two series. Figure 1 shows a rough representation of this matrix. We then evaluated each distance on its capacity to

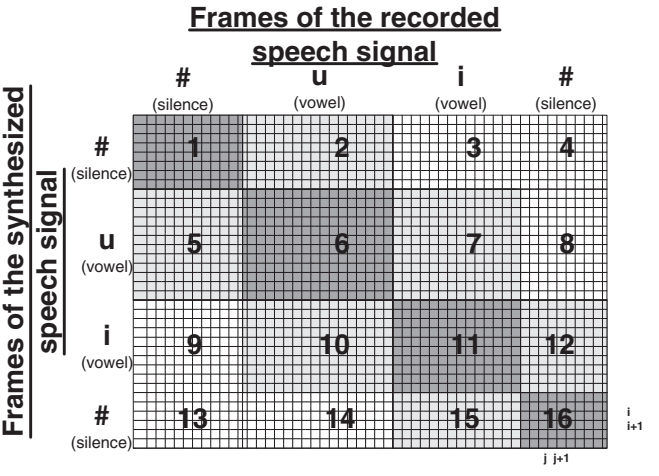

Fig. 1. Graphical representation of the distance matrix regions used for choosing the best feature / pair of features to align the different pairs of phonetic segments

discriminate the difference between two consecutive phones. This was achieved by computing the average distance between feature vectors of the same phone ($\overline{dist}_s$), and of different phones ($\overline{dist}_d$). Using the example in Fig. 1, if we want to choose an acoustic feature to distinguish the *silence* (#) and the vowel *u*, the $\overline{dist}_s$ is the average of the values in regions 1 and 6 on that matrix, while the $\overline{dist}_d$ is the average of the values on regions 2 and 5.

This procedure was performed for every pair of phones and for every utterance on the training set, and its resulting values were saved at the end of each iteration. Finally, we computed an average value of the ratio between $\overline{dist}_s$ and $\overline{dist}_d$ for each pair of phonetic segments and for each acoustic feature. The chosen feature is the one that gives a minimal value for this ratio using the equation:

$$F_k = \min_x \sum_{i=1}^{N_k} \frac{\overline{dist}_s(k,x,i)}{\overline{dist}_d(k,x,i)} \tag{3}$$

where, x is one of the tested features, k represents the pair of phones that is being analyzed, N_k is the number of instances of this pair in our set of utterances, F_k is the best feature for this type of transition, and $\overline{dist}_s(k,x,i)$ and $\overline{dist}_d(k,x,i)$ are the mean distances for the instance i using the acoustic feature x. The smaller is that ratio, the greater is probability of having well aligned frames, locally at least. With this approach, we are trying to use the features that assign the greatest penalty for the alignment paths when they fall out of the darkest regions of Fig. 1 (regions 1, 6, 11 and 16).

Given the reduced amount of training data, we soon realized that it would be impossible to have a large enough number of instances, for each pair of segments to produce confident results. Thus the different phonetic segments were grouped into phonetic classes: vowels, fricatives, plosives, nasals, liquids and silence. The

Table 1. Best feature pairs for the multiple phonetic segment class transitions

	Nasals	Fricatives	Liquids	Plosives	#	Vowels
Nasals	frm+lsf	mfcc+zcrs	frm+en	lsf+en	frm+en	mfcc+mfcc
Fricatives	lsf+lsf	mfcc+en	en+zcrs	lsf+en	zcrs+en	lsf+lsf
Liquids	lsf+en	lsf+en	lsf+lsf	mfcc+en	mfcc+en	frm+mfcc
Plosives	lsf+en	lsf+lsf	lsf+en	mfcc+mfcc	lsf+zcrs	mfcc+en
#	lsf+en	lsf+en	lsf+en	lsf+en	x	lsf+en
Vowels	mfcc+en	zcrs+lsf	mfcc+en	lsf+en	mfcc+en	frm+mfcc

semi-vowels were grouped into the class of the vowels. The described procedure for differentiating the phones was then repeated using phone class transitions (vowel-vowel, fricative-vowel, etc.).

The analysis of the results showed that, in general, for each pair of phone class transition, at least two of the selected features showed good discriminative capacity. This could suggest some equivalence between the two features but it could also mean that the two features were complementary. This way we performed a combined optimization to select the pair of features for each phone class pair. The process could be extended to a combination of even more features but the results showed that there was no significant improvement in using more than a pair of features. The Table 1 shows the results of this procedure, where mfcc, lsf, frm, en and zcrs are the MFCC coefficients and their deltas, LSFs, formants, energy and its delta, and the zero crossing rate, respectively. The x symbol means that this class transition does not exist in the training set. The best feature pair for a transition x-y, is located on the line of x and column of y.

4 Frame Alignment

Before applying the DTW algorithm the distance measure matrix between the reference and the spoken signal must be built. Since we know the boundary locations of the synthetic segments, the distance matrix can be built iteratively, phone-pair by phone-pair.

Taking the example shown in Fig. 1, to build the distance matrix we start by computing the matrix values for all the rows that correspond to the phone-pair #-u using the best pair of features, based on the former results. However, the phone u also belongs to the next phone-pair (u-i) and the computed distance is multiplied by a decreasing triangular weighting window. The distance for the next phone pair (u-i) is then computed using the best pair of features for the vowel-vowel transition and its value is added to the rows corresponding to segment u weighted by an increasing triangular window. Figure 2 shows this weighting triangular windows, where the dotted lines are the weighing factor of the previous phone-pair distances and the dashed lines are the weights of the distances for the next phone-pair. After computing all the values of the distance matrix, the DTW algorithm is applied to find the path that links the top left corner of the matrix to the lower right corner with a minimum accumulated

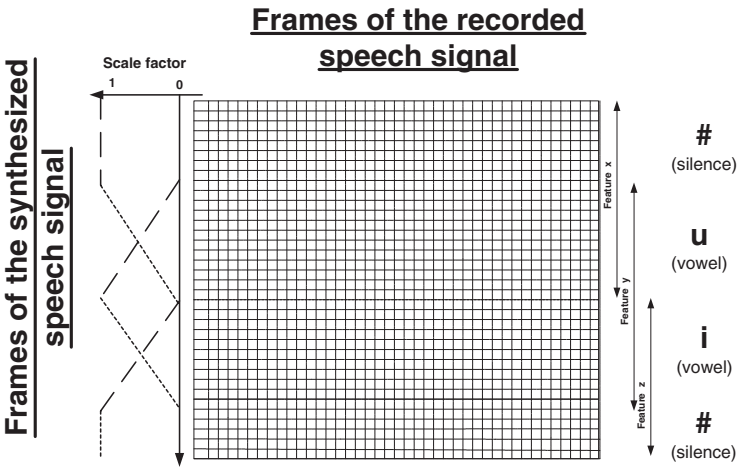

Fig. 2. Graphical representation of the necessary operations for building the distance matrix

distance. This path will be the alignment function between the time scale of the synthetic reference signal and the spoken utterance.

5 Results

The procedure described in the previous section was applied to the reference corpus of 300 manually annotated sentences. The results are depicted in Fig. 3 where the lower solid line is the annotation accuracy when the entire set is aligned using always a feature vector 12 Mel-frequency cepstrum coefficients and their differences. Only 46% of the phonetic segments were aligned with an error less than 20 ms. Using only the best feature for computing the distance for each phone class pair increases the 20ms accuracy to 59% of the segments (dashed line). This result can be improved to 70% by combining two features for computing the distance measure.

The relatively low percentage of agreement for tolerances lower than 20ms can be partially explained by the fact that the segmentation criteria used in the annotation of the reference corpus was not exactly the same as the one used in the segmentation of the logathomes used to produce the synthetic reference. Another difficulty was that the speech material in the reference corpus was uttered by a professional speaker with a very rich prosody and large variations in energy, where several consecutive voiced speech segments become unvoiced. This is, in our opinion the main reason for about 4% of disagreement within high tolerances (about 100 milliseconds). We hope to detect this alignment problems with some confidence measures based on the alignment cost per segment and by phone duration statistics. As soon as we have more annotated material we also plan to

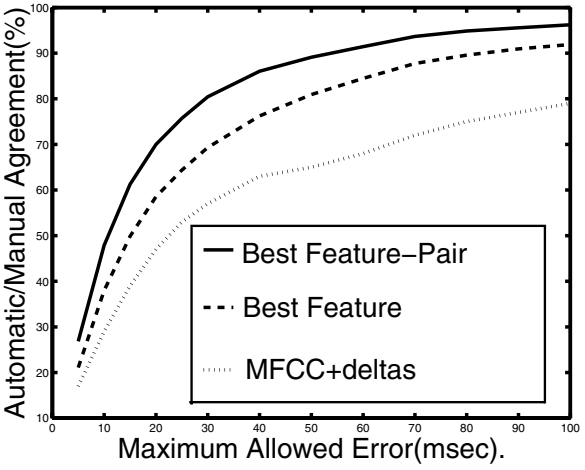

Fig. 3. Accuracy of some versions of the proposed algorithm and a classic DTW-based phonetic alignment algorithm

evaluate the annotation accuracy for a corpus on which we had not optimize the feature selection in order to test the generality of the selected features

6 Conclusions

In this work we have presented a method for selecting the most relevant pair of features for aligning two speech signals with the same phone pairs but with different durations. This features were then used in a DTW-based method for performing the phonetic alignment of a spoken utterance. The results clearly show the advantage of selecting the most appropriate features for each class of segments in the alignment of two utterances: the most commonly used feature, MFCCs, performed well bellow the proposed method.

Acknowledgements. The authors would like to thank M. Céu Viana and H. Moniz for providing the manually aligned reference corpus. This work is part of Sérgio Paulo's PhD Thesis sponsored by a Portuguese Foundation for Science and Technology (FCT) scholarship. INESC-ID Lisboa had support from the POSI Program.

References

1. M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, *The AT&T Next-Gen TTS System*, 137th Acoustical Society of America meeting, Berlin, Germany, 1999.
2. A. Black, *CHATR, Version 0.8, a generic speech synthesizer*, System documentation, ATR-Interpreting Telecomunications Laboratories, Kyoto, Japan, 1996.

3. Sakoe H. and Chiba, *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Trans. on ASSP, 26(1):43–49, 1978.
4. S. Paulo and L. Oliveira, *Multilevel Annotation of Speech Signals Using Weighted Finite State Transducers.* In Proceedings of IEEE 2002 Workshop on Speech Synthesis, Santa Monica, California, 2002.
5. D. Caseiro, H. Meinedo, A. Serralheiro, I. Trancoso and J. Neto, *Spoken Book alignment using WFST* HLT 2002 Human Language Technology Conference, San Diego, California, 2002.
6. F. Malfrère and T. Dutoit, *High-Quality Speech Synthesis for Phonetic Speech Segmentation.* In Proceedings of Eurospeech'97, Rhodes, Greece, 1997.
7. N. Campbell, *Autolabelling Japanese TOBI.* In Proceedings of ICSLP'96, Philadelphia, USA, 1996.
8. A. Black, P. Taylor and R. Caley, *The Festival Speech Synthesis System.* System documentation Edition 1.4, for Festival Version 1.4.0, 17th June 1999.
9. P. Taylor R. Caley, A. Black, S. King, *Edinburgh Speech Tools Library* System Documentation Edition 1.2, 15th June 1999.
10. *ESPS Programs Version 5.3* Entropic Research Laboratories Inc., 1998.