

Phonetic Alignment : Speech Synthesis-based vs. Viterbi-based.

F.Malfrère^{1,2}, O. Deroo^{1,2}, T. Dutoit¹ and C. Ris¹

¹Faculté Polytechnique de Mons — TCTS

31, Bld. Dolez

B-7000 Mons, Belgium

Email: malfrere,deroo,dutoit,ris@tcts.fpms.ac.be

²Now with Babel Technologies SA

Boulevard Dolez 33, 7000 - Mons (Belgium)

tel.: +32.65.37.42.78

Email: malfrere@babeltech.com

19th February 2002

Abstract

In this paper we compare two different methods for automatically phonetically labeling a continuous speech database, as usually required for designing a speech recognition or speech synthesis system. The first method is based on temporal alignment of speech on a synthetic speech pattern; the second method uses either a continuous density HMM (Hidden Markov Models) or a hybrid HMM/ANN (Artificial Neural Network) system in forced alignment mode. Both systems have been evaluated on read utterances not part of the training set of the HMM systems, and compared to manual segmentation. This study outlines the advantages

and drawbacks of both methods. The speech synthetic system has the great advantage that no training stage (hence no large labeled database) is needed, while HMM systems easily handle multiple phonetic transcriptions (phonetic lattice). We deduce a method for the automatic creation of large phonetically labeled speech databases, based on using the synthetic speech segmentation tool to bootstrap the training process of either a HMM or a hybrid HMM/ANN system. The importance of such segmentation tools is a key point for the development of improved multilingual speech synthesis and recognition systems.

Résumé

Nous comparons dans ce papier deux méthodes pouvant être utilisées pour annoter phonétiquement et de manière automatique un corpus de parole continue, comme c'est généralement nécessaire pour la mise au point de systèmes de reconnaissance ou de synthèse de la parole. La première méthode est basée sur un alignement temporel du signal de parole sur un signal synthétique de bonne qualité. La seconde méthode utilise soit des modèles de Markov cachés HMM à distributions multigaussiennes ou un système hybride HMM/ANN en mode d'alignement forcé. Les deux systèmes ont été évalués sur des phrases lues n'ayant pas servi à l'entraînement des systèmes (HMM ou hybride) et manuellement segmentées. Cette étude met en évidence les avantages et inconvénients de chacune des méthodes. Le système basé sur le synthétiseur a le grand avantage qu'aucune phase d'entraînement (et donc aucun grand corpus segmenté) n'est nécessaire, alors que les systèmes classiques basés sur les HMMs peuvent facilement prendre en compte des transcriptions phonétiques multiples. Nous avons ainsi mis au point une méthode pouvant être utilisée pour la création automatique de corpus de parole phonétiquement étiquetés. Cette méthode est basée sur l'utilisation du système d'alignement basé sur le signal synthétique afin d'initialiser le processus d'entraînement d'un système HMM (gaussien ou hybride). Ces méthodes de segmentation automatique sont d'une grande importance pour le développement de

systèmes de synthèse et de reconnaissance de la parole multilingues.

Keywords : *Speech Segmentation, Hidden Markov Models, Hybrid HMM/ANN systems, Speech Synthesis, Large Speech Corpora.*

Contents

1	Introduction	7
2	The segmentation problem	8
3	The HMM-based method	9
3.1	HMM models	9
3.2	Hybrid HMM/ANN systems	9
4	Speech Synthesis-Based Phonetic Alignment	11
5	Experiments	13
5.1	The American English hybrid HMM/ANN system	14
5.2	The French hybrid HMM/ANN system	14
5.3	The Dutch hybrid HMM/ANN system	15
5.4	The Spanish hybrid HMM/ANN system	16
6	Results	16
6.1	American English : TIMIT	17
6.2	French : BDSONS	17
6.3	Dutch : COGEN	19
6.4	Spanish : LATINO-40	19
6.5	Results Analysis	20
7	Conclusions	21

List of Figures

1	A hybrid HMM/ANN speech segmentation system.	23
2	A speech synthesis-based alignment system.	24
3	Local continuity condition.	25

List of Tables

1	<i>An example of local continuity constraints expressed in terms of coordinate increments for the DTW process.</i>	26
2	<i>Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the US-English database.</i>	27
3	<i>Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the French database.</i>	28
4	<i>Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the Dutch database.</i>	29
5	<i>Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the Spanish database.</i>	30
6	<i>Segmentation accuracy on the TIMIT database using the MBROLIGN, HMM and Hybrid HMM/ANN-based methods.</i>	31
7	<i>Segmentation accuracy on the BDSONS database using the MBROLIGN and Hybrid HMM/ANN-based methods. Third set of results corresponds to speaker normalization applied before the MBROLIGN method</i>	32
8	<i>Segmentation accuracy on the COGEN database using the MBROLIGN and Hybrid HMM/ANN-based methods.</i>	33
9	<i>Segmentation accuracy on the LATINO-40 database using the MBROLIGN and Hybrid HMM/ANN methods.</i>	34

1 Introduction

This paper focuses on text-to-speech alignment – that is, on the alignment of a phonetic transcription with the corresponding speech signal. TTS alignment tools have become a key point for the development of very large segmented (and annotated) speech databases, which are now required in all areas of speech technology. For speech synthesizers based on diphone concatenation, a complete set of diphones must be extracted from a set of predefined words. Manual diphone segmentation is a time consuming and tedious operation. The automatization of this task allows a faster creation of new voices for diphone-based speech synthesizers. Moreover, the development of the new generation of speech synthesizers (Hunt and Black 1996) based on non-uniform unit selection requires single speaker speech corpora of several hours entirely phonetically annotated. Speech recognition systems (Baker 1975, Russel et al. 90, Bahl et al. 1995, Rabiner and Juang 1993, Woodland et al. 1995) also require several hours of multispeaker speech data. Last but not least, prosody generation systems are also great consumers of large annotated speech corpora (Traber 1995, Malfrère and Dutoit 1997).

Corpus-based methods tend to have the final word over models based on human expertise. Text-to-speech alignment, however, is still often done by hand, or at least corrected by an expert. To reduce the costs, automatic segmentation tools are required. Ideally, such systems must be speaker-independent, language-independent and must provide an accurate and consistent segmentation. Manual segmentation is speaker-independent and provides accurate but not consistent nor reproducible segmentation (Cosi et al. 1991). Automatic speech segmentation provides a consistent and reproducible segmentation, but is more error prone. The literature reports many HMM-based automatic labeling systems (Leung 1984, Ljolje and Riley 1991, Brugnara et al. 1993, Talkin and Wightman 1996), most of them evaluated on one language only (English most of the time). Only the system proposed by Vostermans, Martens and

Van Coile (Van Coile et al. 1994) reports results on American English, French, Spanish, Dutch and German. All these automatic systems require a training stage for which large phonetically labeled speech corpora are needed. In this paper, we study two methods to perform text-to-speech alignment. The first method is based on the classical HMM approach and is described in Section 2. The second approach is based on the use of a speech synthesizer to create a reference signal on which natural speech can be aligned (Deroo et al. 1998, Lenzo and Black 2000, Horak 2001). This method does not require any training stage and is described in section 3. In section 5, we describe the training of HMM systems for several languages and give recognition rates. Section 6 reports the results obtained with both approaches on American English, Dutch, Spanish and French.

2 The segmentation problem

The purpose is to establish a one-to-one correspondence between a sequence of contiguous speech segments and a sequence of phonetic labels. As usual in speech processing, the sampled speech waveform is not used directly. Each sentence is represented by a sequence of acoustic vectors characterizing the speech signal over a small time frame of typically 10 to 30 ms with successive frame shifts of 10ms. Thus the segment boundaries can only be expressed in terms of number of frames, and the accuracy is inherently limited by the frame shift. Many automatic labeling systems have already been reported in the literature; all of them use Hidden Markov Models (HMM) (Leung 1984, Ljolje and Riley 1991, Brugnara et al. 1993, Talkin and Wightman 1996).

As a first stage of the segmentation process, the phonetic transcription of the sentences to segment must be obtained. In our case, they are automatically derived from the text with an accurate automatic phonetization system used in text-to-speech synthesis systems.

3 The HMM-based method

3.1 HMM models

HMMs are now widely used in speech recognition (Baum 1972, Jelinek 1976, Myers and Rabiner 1981, Rabiner and Juang 1993). HMMs are able to take the statistical variability of speech into account. The training procedure is well known. In all the experiments reported here, we used embedded Viterbi training, a procedure in which the phonetic labeling of the database is recursively refined using forced Viterbi alignment. The segmentation process of a speech database require Context Independent phoneme-based HMM models. The parameters are initialized using manually segmented speech material or applying linear initialization. A linear segmentation consists in assigning to each phoneme of a sentence a number of frames which is equal to the total number of frames in the sentence divided by the number of phonemes in the sentence (linear), or proportional to the average duration of each phoneme (linearly weighed), eventually assisted by a speech/silence detector. Once these initial parameters have been obtained, supervised Viterbi training is applied, which results in a new segmentation from which it is possible to update the HMM parameters. The main problem of this method is that a first segmentation is required to bootstrap the training process. Linear segmentation can be used for isolated word databases. But the training process based on linear segmentation leads to many convergence problems for long sentences. That is why HMM parameters are usually initialized using (at least partially) hand-labelled databases.

3.2 Hybrid HMM/ANN systems

The HMM systems are mostly based on gaussian mixture models (GMM) which are used to modelize the state emission probability density functions. An alternative approach are the hybrid HMM/ANN systems (Boulevard and Morgan 1994, Robinson and Fallside 1991, Robinson 1994) which combine the advantages of HMMs and those of Artificial Neural Networks :

1. They provide discriminant learning.
2. When used in classification mode and trained with a Least Mean Square criterion or with an entropy criterion, the network outputs are estimates of posterior probabilities. This is achieved without requiring strong assumptions about the underlying probability density functions.
3. ANNs can make use of contextual information by taking multiple frames as input.

Hybrid HMM/ANN models have already proved their ability to obtain very good performance on many different tasks (from speaker-independent, medium vocabulary, isolated word recognition (Dupont et al. 1997) to large vocabulary continuous speech recognition (Franco et al. 1994, Hochberg et al. 1995)). As the training of ANNs is more time-consuming than that of GMM-based HMMs, accurate segmentation is needed to initialize the training process. This guarantees a quicker convergence to accurate acoustic models. Additionally if the segmentation used to bootstrap the training is not good enough (linear for instance), convergence problems to a local minimum may occur. Context-independent phoneme GMM-based HMMs can be used for generating a first segmentation used to bootstrap the hybrid HMM/ANN system. But even in this case, we still have the problem of initializing the HMM parameters and many efforts and time are lost in improving the baseline segmentation. In all the experiments reported in this paper, we used a HMM/MLP (Multi-layer Perceptron) hybrid model (see figure 1 for the complete alignment system) as well as HMMs, for comparison with the method developed in section 4. A minimum duration of half the average duration of each phoneme was used to define the Context Independent model topologies.

Figure 1 should appear here.

4 Speech Synthesis-Based Phonetic Alignment

The main idea of the speech synthesis-based phonetic alignment is to use a speech synthesizer to create a reference speech pattern with predetermined phonetic segmentation and then align natural speech on this pattern (figure 2). The publicly available MBROLA (Dutoit et al. 1996) speech synthesizer, which is based on diphone concatenation, is used to generate a reference synthetic speech signal from the phonetic transcription of the sentences. Although natural prosodic information is needed to deliver natural sounding synthetic speech, a very rough prosody suffices to obtain the reference signal since only its segmental features will be used during the temporal alignment process. Phoneme duration and intonation contours are thus chosen so as to facilitate the alignment process. A constant duration of 100 ms has been chosen to synthesize all the phonemes.

Figure 2

Since no assumption can be made on the contour actually produced by the speaker, the synthetic F0 curve is chosen as simple as possible (a constant F0 value). Assuming the features used to compare the reference and the test signals are not correlated with the F0 curve, this choice has no important effect on the accuracy of the segmentation.

should appear here.

To compare the synthetic reference speech and the original speech, some relevant features must be extracted from both signals. Four sets of parameters have been used to characterize speech frames:

- The 12 first cepstral coefficients (c_i) derived from a linear prediction analysis (10th order). These coefficients are normalized (Cepstral Mean Subtraction CMS) and weighted with a sinusoidal function (Juang et al. 1986).
- Delta cepstral coefficients (Δc_i) in order to account for speech dynamics.
- The normalized energy (E) of each frame.
- Its delta energy (ΔE).

The resulting 26 coefficients are known to result in a good representation of the local spectral envelope.

Finally, the segmentation process takes place. It is based on a classical dynamic time warping (DTW) algorithm based on the minimization of the accumulated distance between the two speech signals. The distance used to compare a frame a of the synthetic reference and a frame b of the input speech is a weighted combination of several euclidian distances: the cepstral and Δ cepstral distances and the energy and Δ energy distances (see Equation (1)).

$$d(a, b)^2 = \alpha \sum_{j=0}^N (c_j(a) - c_j(b))^2 + \beta \sum_{j=0}^N (\Delta c_j(a) - \Delta c_j(b))^2 + \gamma (E(a) - E(b))^2 + \varphi (\Delta E(a) - \Delta E(b))^2 \quad (1)$$

The optimization of the orders and of the LPC and cepstral analysis weighting coefficients α , β , γ , φ has lead to the following parameters:

- Frame of 30 ms with a overlap of 20 ms and a shift of 10 ms
- Linear prediction order: 10 (sampling rate = 16 KHz)
- Cepstral analysis order: $N = 12$
- $\alpha=1.0; \beta=1.25; \gamma=1.25; \varphi=1.25$

A constant phoneme duration of a hundred milliseconds has been chosen for the reference synthetis signal (Malfrère and Dutoit 1997). To ensure proper time alignment a local continuity constraint is used as in (Rabiner and Juang 1993). Its form is represented on Figure 3. The constraint expresses the allowable paths to reach a given point in the grid mapping the original signal on the synthetic reference signal. Each allowable path is defined as a sequence of moves, each of which is specified by a pair of coordinate increments on the grid (Figure 3 and Table 1 illustrates 15 paths).

The great advantage of this approach is that there is no training stage, and so no training database is needed. As a result, the system can be easily adapted

*Figure 3
and Table 1 should
appear here.*

to align different languages provided a speech synthesizer is available for it (which is now the case for 24 languages in the MBROLA project). Segmentation results are given in Section 6 for English, German, Dutch, French, Spanish and in (Malfrère and Dutoit 1997) for Romanian.

One of the drawbacks of alignment on a synthetic voice is the speaker dependency of the system. Indeed, the same reference voice is used for every alignment whoever the speaker is. This effect could be reduced by applying some speaker normalization. However, results given in the next sections show that this effect is not of prime importance (see section 6.2).

This system has been integrated in an interactive prosody transplantation tool called MBROLIGN which can be freely downloaded for academic purposes from our web site : <http://tcts.fpms.ac.be/synthesis/mbrolign>.

5 Experiments

In this section we describe how we trained the hybrid HMM/ANN system for each language examined. For simplicity, we used the SAMPA (SAM Phonetic Alphabet definition, ESPRIT Project 2589, 1992) phone set to annotate the phonemes used in each language (for the training and test databases). This gave 35 phonemes for French, 42 phonemes for US English, 24 for Spanish and 45 for Dutch.

Two sets of acoustic features have been used: the Perceptual Linear Predictive coefficients (PLP) (Hermansky 1990) and the log-RASTA-PLP coefficients (Koehler et al. 1994). These parameters have been chosen for their robustness against channel and speaker characteristics. They were computed every 10 ms over 30 ms analysis windows. The order of the LPC analysis was set to 10.¹

¹We compared different kind of parameters and we observed that the best parameters for the MROLIGN-based system were the cepstral coefficient with Cepstral Mean Substraction. For the hybrid HMM/ANN system the best results were obtained with RASTA-PLP or PLP coefficient. The difference between both parameters Cepstral vs Rasta was not very important but in all the experiment reported here, we only give the best results obtained on the different databases.

The feature set for the hybrid HMM/ANN systems was a 26-dimensional vector composed of the *cepstral* parameters (PLP or log-RASTA-PLP), the Δ *cepstral* parameters, the Δ *energy* and the $\Delta\Delta$ *energy*. Nine frames of contextual information were used at the input of the ANNs, leading to 234 inputs (9 frames of context being known as yielding usually the best recognition performance (Bouclard and Morgan 1994)). In order to have an overview of how well the Neural Networks are trained, we give the training and cross validation rate for each language. Those scores give the percentage of correctly classified frames for the training set and for a validation set (the validation set is commonly used in Neural Networks training in order to stop the training phase before overtraining) (Bouclard and Morgan 1994).

5.1 The American English hybrid HMM/ANN system

As for the French model, we used a large read speech corpus WSJ0 (Paul and Baker 1992) with text material selected from the *Wall Street Journal* newspaper so as to provide a representative range of phonetic environments. The corpus was split into speaker-dependent and speaker-independent subsets, and further split into 5,000-word and 20,000-word vocabularies.

The official WSJ0 database training set has been used in order to train our acoustic models. This consists of approximately 12000 utterances pronounced by 112 speakers. The training and cross-validation scores at the frame level are given in table 2.

*Table 2
should appear here.*

5.2 The French hybrid HMM/ANN system

We used the BREF-80 (Lamel et al. 1991) database in order to train the French hybrid HMM/ANN system. BREF-80 is a large read speech corpus with 80 speakers. The text material was selected from the French newspaper *Le Monde* so as to provide large vocabulary coverage (over 20.000 words) and a representative range of phonetic environments. As BREF contains 1115 distinct diphones and over 17.500 triphones, it can be efficiently used to train phonetic

models. The phonetic transcriptions of these texts were obtained using a text-to-phoneme tool. The training set used in the following experiments consists of 3737 sentences (3363 sentences for training and 374 for cross validation²) from 56 speakers (approximately 9 hours of speech).

As no phonetic segmentation is provided with BREF, we generated a first segmentation using the bootstrapping method (Malfrère and Dutoit 1997) described in section 4, and then iterated the training process.

The training and cross-validation scores on this particular database at the frame level are given in table 3.

*Table 3
should appear here.*

5.3 The Dutch hybrid HMM/ANN system

This system has been trained on a particular database recorded at FUNDP (Facultés Universitaires Notre Dame de la Paix, Namur Belgian) for the DEMOSTHENES project (Deville et al. 1999). The DEMOSTHENES database consists of isolated words, phrases and sentences that are representative of the pronunciation errors made by French-speaking learners (e.g. language-specific phonemes without equivalent in French, assimilations, confusion between long/short vowels, etc.). About 25 different pronunciation difficulties are illustrated in a sample of several hundred items, pronounced by 135 (native and non-native) speakers of Dutch. The phrases and sentences of the database have been carefully selected so as to cover the basic vocabulary of Dutch (2000 most frequent words) to provide a representative range of phonetic environments. Only the native speakers have been used to train the system. We are thus sure that no pronunciation errors have been used to train the acoustic models. The training and cross-validation scores at the frame level are given in table 4.

*Table 4
should appear here.*

²Used to adapt the learning rate of the MLP (Bourlard and Morgan 1994).

5.4 The Spanish hybrid HMM/ANN system

The hybrid HMM/ANN Spanish system has been trained on the LATINO-40 database³. This database provides a set of recordings for training speaker-independent systems to recognize Latin-American Spanish. The database comprises about 5000 utterance files : about 125 utterances (apparently from Latin American newspaper texts) from each of 40 different speakers, 20 males and 20 females. The sentences are all shorter than 80 characters, and are not grouped into larger constituents like paragraphs or stories.

The training set was composed of 4200 utterances (4000 for training and 200 for cross validation). The training and cross-validation scores at the frame level are given in table 5.

*Table 5
should appear here.*

6 Results

The following databases have been used in order to compare the method developed in section 4 with the Gaussian HMM and hybrid HMM/ANN systems in section 3 : TIMIT for American English, BDBSONS for French, COGEN for Dutch, LATINO-40 for Spanish. All databases are 16 bits/ 16 KHz. They are either provided with hand labeling or have been carefully hand segmented and manually checked by a single expert in our laboratory to generate a reference segmentation.

For each language we will give a table with the segmentation accuracy with all the different methods explained above for four broad phonetic classes (VV : Vowel-Vowel; VC : Vowel-Consonnant, CC Consonnant-Consonnant and CV Consonnant-Vowel), as a function of the time shift between the segmentation obtained and the reference segmentation (lower than 10ms, 20 ms, 30 ms, 40 ms, 50 ms and greater than 50 ms). The classes we have chosen are a compromise between a single class (hence an overall alignment score, which provided poor analytical information) and phoneme-by-phoneme results (which would be hard

³Distributed by LDC : <http://www.ldc.upenn.edu/>.

to read and compare).

6.1 American English : TIMIT

The TIMIT (Zue et al. 1990) corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains speech from 630 speakers from 8 major dialects of American English, each speaking 10 phonetically rich sentences. This corpus includes hand made time-aligned orthographic, phonetic, and word transcriptions for each sentence. Table 6 gives the results obtained with the 3 different methods. The HMM multi-gaussians system has been trained on the same database as the hybrid HMM/ANN system⁴. The multi-gaussians system used full covariance matrices and the number of Gaussians per state was set to 16⁵.

*Table 6
should appear here.*

The hybrid HMM/ANN system clearly outperforms the HMM for each category (except for the VV where the HMM is slightly better) and provides more accurate segmentation. This can be explained by the properties of the hybrid HMM/ANN systems, which are known to be much more accurate at the phoneme level than HMMs because of their discriminant behaviour. The segmentation obtained with MBROLIGN is slightly less accurate than that obtained with the other systems for VV transitions. It is better or equivalent to hybrid otherwise. In the following experiments only the hybrid HMM/ANN system will be compared with the method developed in section 4.

6.2 French : BDSOONS

French alignment was based on the French database BDSOONS (Carré et al. 1984). As this database was not manually labeled, we only selected a part of the database and a single expert labeled it to generate the reference segmenta-

⁴see section 5.1

⁵The number of parameters is approximately the same as the hybrid HMM/ANN system.

tion file. Table 7 gives the results obtained with the MBROLIGN and hybrid HMM/ANN systems. The segmentation obtained with MBROLIGN is slightly less accurate than that obtained with the hybrid system. The segmentation rates are comparable to those obtained for American English.

*Table 7
should appear here.*

In an effort to reduce the apparent loss of speaker indendency inherent to the speech synthesis based method, we propose to apply some speaker normalization (Lee and Rose 1998), namely the vocal tract length normalization based on a frequency warping approach. The idea of this approach is to reduce the variability of formant center frequencies due to vocal tract shape variations between speakers by linearly warping the frequency axis. Practically, the warping procedure consists in modifying the Mel filter bank (center frequencies and bandwidth) according to some linear transformation: $G(f) = \alpha \cdot f$.

There are different ways to optimize the warping parameter α . We chose the maximum-likelihood criterion. In this case, the speech signal is processed with a set of 13 discrete values of α ranging from 0.88 to 1.12 and the alignment is performed for each of these values. The optimal warping factor is the one leading to the best alignment score (DTW score) over all the sentences pronounced by a speaker: $\hat{\alpha} = \operatorname{argmax}_{\alpha} P(X^{\alpha}, \lambda)$ ⁶. Once the optimal warping factor has been computed for a given speaker, we process all the sentences with this speaker dependent warping factor and re-align this "normalized" speech with the synthetic voice.

Results of this experiment are appended to Table 7. As can be seen in this table, speaker normalization does not improve significantly the segmentation accuracy. We understand this effect as follows: the speaker normalization process is based on the minimization of the DTW alignment score while the scores given in Table 7 are related to segmentation accuracy. Optimizing the first does not automatically increase the second. Other speaker normalization techniques could of course be investigated but this is beyond the scope of this paper.

⁶ λ represent all the Models.

6.3 Dutch : COGEN

The COGEN (Corpus Gesproken Nederlands) database⁷ contains continuous speech recorded in an anechoic chamber. The sentences were selected from a set of 130 phonetically balanced sentences. This database contains the phonetic labels and their positions, provided by a human expert who performed an audio-visual inspection of the utterance. Hand labeling was performed starting from automatic labeling files obtained with HMMs as the initial estimate. The same expert checked the segmentation of the database to generate the reference segmentation file which was used to compare segmentation methods. Table 8 gives the results obtained on this particular language with the Hybrid HMM/ANN system and the MBROLIGN-based system. Considering the < 20 ms error rate as a reference, the segmentation obtained with MBROLIGN is about 40% less accurate, the segmentation obtained with MBROLIGN is slightly less accurate than that obtained with the hybrid system. The segmentation scores are comparable to those obtained for the other languages.

*Table 8
should appear here.*

6.4 Spanish : LATINO-40

The LATINO 40 database has been used to compare both systems. Here again, a single expert labeled a part of this database (different from the one used in the training of the hybrid HMM/ANN system) to generate the reference segmentation file. The results obtained on this particular language can be found in table 9. Considering the < 20 ms error rate as a reference, the segmentation obtained with MBROLIGN is about 40% less accurate, the segmentation obtained with MBROLIGN is slightly less accurate than that obtained with the hybrid system. The segmentation scores are comparable to those obtained for the other languages.

*Table 9
should appear here.*

⁷developed by the Katholieke Universiteit LEUVEN (KUL) and the University of GENT.

6.5 Results Analysis

The significance of our results (Tables 6, 7, 8, and 9) depends on the application for which the alignment system is being used. If alignment is targeted by itself (for creating a database which will be used to derive statistics on phoneme contextual durations for speech synthesis, for instance), then the reference acceptable error should be that of humans : around 20 ms according to (Cosi et al. 1991). If, on the contrary, alignment is used for prosody transplantation (i.e., copy synthesis using alignment and intonation from the speech data directly as input for a speech synthesis system), then misalignments of even less than 20 ms can lead the synthesizer to produce speech with wrong pitch (typically, pitch computed on frames erroneously considered as part of voiced phonemes). For the following discussion we will use 20 ms as a reference acceptable error.

We observe that the results obtained with the MBROLIGN-based method are slightly better than those obtained with the best HMM system (hybrid HMM/ANN system) for American English. For French, they are slightly worse. For Dutch and Spanish, they are significantly worse (about 40 % more errors). In all cases, the MBROLIGN-based method gives less accurate VV segmentation (for American English, this has little consequence, given the relatively small amount of VV transitions). It should be emphasized however that the results obtained with the MBROLIGN-based method obviously depend on the synthetic voice being used, and cannot be interpreted in favor or against the described method. It follows, from an engineering perspective, that both methods (MBROLIGN-based and HMM-based) are generally comparable. It is therefore not surprising that other similar DTW-based methods have been developed recently (inside the Festival TTS system (Lenzo and Black 2000) and in Speech Studio (Horak 2001)).

This rises another idea : that of using this method to generate a first segmentation that can be used for bootstrapping the HMM training process. In this case, the difference in accuracy between the DTW-based segmentation and state-of-the-art (HMM-based) segmentation is still less important, since DTW-

based segmentation competes with linear segmentation (as often performed as the first alignment approximation) and alignment will be ultimately refined by the EM training algorithm. For training a speech recognition system, we can use the MBROLIGN-based method for the automatic creation of a first segmentation and then use it to bootstrap the training process of continuous densities HMM or hybrid HMM/ANN systems.

7 Conclusions

We compared in this paper two different methods for automatically phonetically labeling a continuous speech database. The first method is based on temporal alignment of speech on a synthetic speech pattern. This method does not require any training stage and so no training database is needed. The system can be easily adapted to align different languages provided a speech synthesizer is available for it (this is the case for 24 languages in the framework of the MBROLA project). The second method is based on the classical HMM approach. They are able to take the statistical variability of speech into account and are now widely used in the speech community. The main problem of this second method is that either a first segmentation is required to bootstrap the HMM training process or already trained models are needed to realize the alignment. We compared HMM with Gaussian Mixture States and Hybrid HMM/ANN systems to the speech synthesis system for several different languages (French, US English, Spanish, Dutch). We observed that the results obtained with the MBROLIGN-based method are comparable or slightly inferior (depending on the language and the speaker) to those obtain with the best HMM system (the hybrid HMM/ANN system). We also tried to decrease the influence of speakers in the MBROLIGN-based approach by performing a Vocal Track Normalization in the signal space. Some experiments in French showed that no improvement in the accuracy of the segmentation system was observed. We conclude with the idea that the MBROLIGN-based system can be used for bootstrapping the HMM training

process of continuous densities HMM or hybrid HMM/ANN systems when no segmentation is available or no pre-trained models in that particular language can be used.

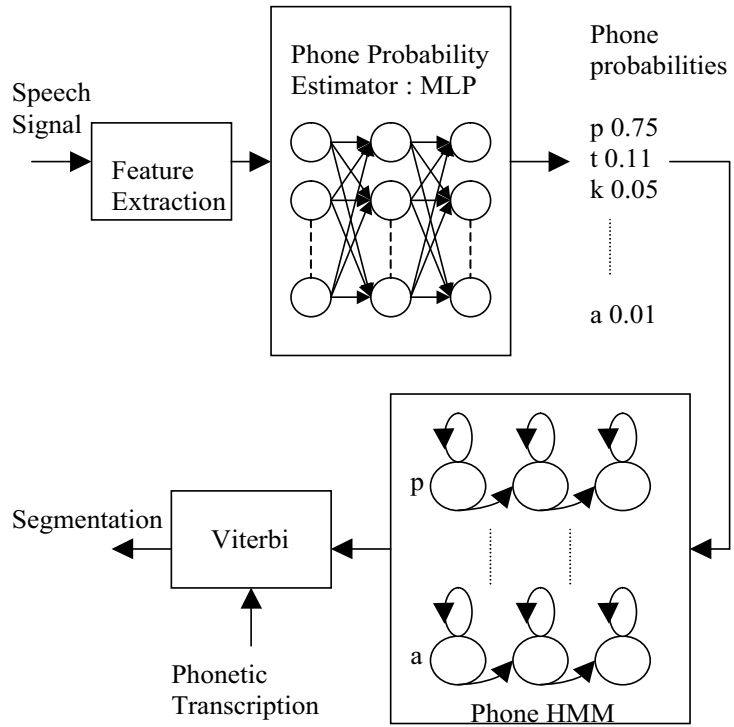


Figure 1: A hybrid HMM/ANN speech segmentation system.

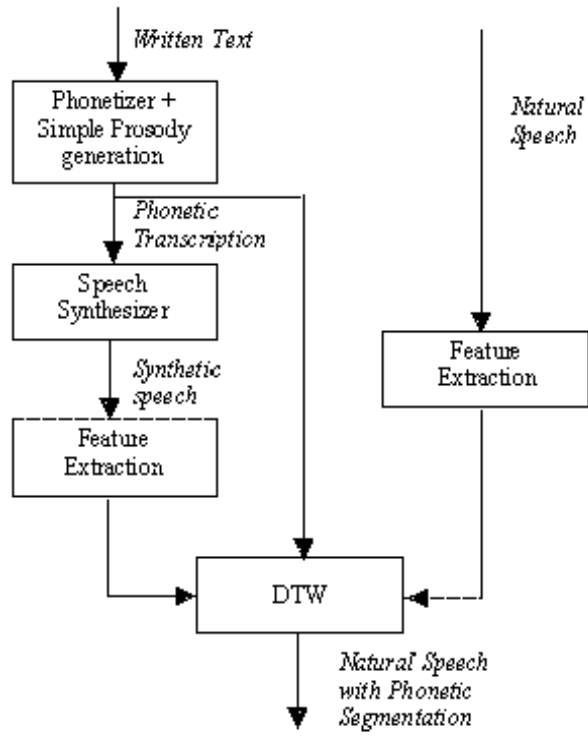


Figure 2: A speech synthesis-based alignment system.

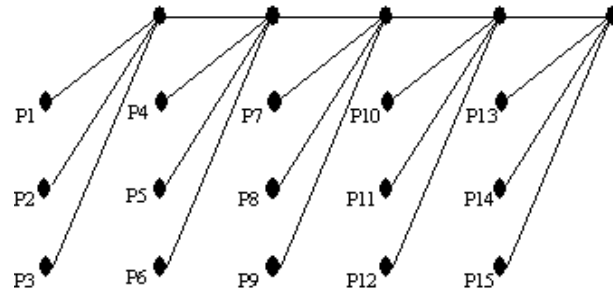


Figure 3: Local continuity condition.

P1: (1,1)(1,0)(1,0)(1,0)	P6: (1,3)(1,0)(1,0)(1,0)	P11: (1,2)(1,0)
P2: (1,2)(1,0)(1,0)(1,0)	P7: (1,1)(1,0)(1,0)	P12: (1,3)(1,0)
P3: (1,3)(1,0)(1,0)(1,0)	P8: (1,2)(1,0)(1,0)	P13: (1,1)
P4: (1,1)(1,0)(1,0)(1,0)	P9: (1,3)(1,0)(1,0)	P14: (1,2)
P5: (1,2)(1,0)(1,0)(1,0)	P10: (1,1)(1,0)	P15: (1,3)

Table 1: *An example of local continuity constraints expressed in terms of coordinate increments for the DTW process.*

	Frames	log-RASTA	PLP
Train	7,878,000	74.1%	76.8%
Cross	880,000	73.8%	75.9%

Table 2: *Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the US-English database.*

	Frames	log-RASTA	PLP
Train	2,400,000	78.6%	79.9%
Cross	270,000	76.0%	77.4%

Table 3: *Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the French database.*

	Frames	log-RASTA	PLP
Train	1,379,000	80.2%	82.8%
Cross	141,000	76.5%	78.9%

Table 4: *Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the Dutch database.*

	Frames	log-RASTA	PLP
Train	1,824,906	83.6%	85.7%
Cross	201,635	82.6%	85.3%

Table 5: *Recognition rate at the frame level using a hybrid HMM/ANN system trained on PLP and log-RASTA-PLP coefficients for the Spanish database.*

Nb Instances	2277	49810	38713	49976	
MBROLIGN	VV	VC	CC	CV	Total
<10 ms	35.97%	58.59%	57.69%	51.12%	55.32%
<20 ms	58.57%	79.60%	80.32%	81.73%	80.21%
<30 ms	72.13%	87.98%	90.55%	90.54%	89.34%
<40 ms	81.05%	91.96%	94.73%	94.27%	93.36%
<50 ms	87.89%	94.46%	96.89%	96.51%	95.75%
>50 ms	12.11%	5.54%	3.11%	3.49%	4.25%
HMM					
<10 ms	51.22%	44.48%	43.86%	44.80%	44.53%
<20 ms	70.92%	71.93%	71.99%	72.01%	71.96%
<30 ms	84.69%	86.82%	85.90%	86.78%	86.52%
<40 ms	92.35%	92.65%	93.16%	92.55%	92.75%
<50 ms	95.51%	95.90%	96.79%	95.78%	96.09%
>50 ms	4.49%	4.10%	3.21%	4.22%	3.91%
Hybrid					
<10 ms	42.52%	52.53%	58.83%	55.90%	55.29%
<20 ms	69.04%	76.43%	82.52%	77.00%	78.18%
<30 ms	83.17%	89.09%	92.42%	88.47%	89.67%
<40 ms	91.71%	94.67%	96.75%	93.41%	94.75%
<50 ms	95.25%	97.33%	98.55%	98.15%	97.93%
>50 ms	4.75%	2.67%	1.45%	1.85%	2.07%

Table 6: Segmentation accuracy on the TIMIT database using the MBROLIGN, HMM and Hybrid HMM/ANN-based methods.

Nb Instances	96	2554	756	2416	
MBROLIGN	VV	VC	CC	CV	Total
<10 ms	50.00%	68.95%	66.43%	69.62%	68.58%
<20 ms	70.00%	82.51%	82.78%	81.98%	82.11%
<30 ms	78.00%	87.86%	89.10%	87.79%	87.82%
<40 ms	86.00%	92.41%	93.50%	93.02%	92.70%
<50 ms	92.00%	95.39%	96.31%	95.06%	95.32%
>50 ms	8.00%	4.61%	3.69%	4.94%	4.68%
Hybrid					
<10 ms	84.73%	66.67%	81.05%	79.12%	80.49%
<20 ms	87.93%	70.59%	83.89%	83.36%	83.97%
<30 ms	91.87%	82.35%	88.85%	89.37%	89.35%
<40 ms	95.07%	86.27%	92.35%	92.53%	92.68%
<50 ms	97.04%	92.16%	95.34%	95.30%	95.49%
>50 ms	2.96%	7.84%	4.66%	4.70%	4.51%
MBROLIGN NORMALIZED					
<10 ms	62.17%	54.17%	59.67%	68.09%	63.39%
<20 ms	76.85%	62.71%	71.93%	77.90%	74.89%
<30 ms	86.24%	75.00%	84.49%	83.36%	84.09%
<40 ms	93.25%	88.54%	92.87%	94.37%	93.47%
<50 ms	96.56%	92.71%	95.14%	95.53%	95.45%
>50 ms	3.44%	7.29%	4.86%	4.47%	4.55%

Table 7: *Segmentation accuracy on the BDBSONS database using the MBROLIGN and Hybrid HMM/ANN-based methods. Third set of results corresponds to speaker normalization applied before the MBROLIGN method*

Nb Instaces	1593	3141	1904	3274	
MBROLIGN	VV	VC	CC	CV	Total
<10 ms	60.81%	70.37%	57.14%	76.19%	68.21%
<20 ms	75.68%	76.30%	73.47%	81.63%	77.42%
<30 ms	85.14%	87.41%	87.76%	89.80%	87.90%
<40 ms	89.19%	91.85%	94.90%	93.20%	92.45%
<50 ms	94.59%	96.30%	95.92%	96.60%	96.05%
>50 ms	5.41%	3.7%	4.08%	3.4%	3.95%
Hybrid					
<10 ms	66.48%	73.09%	51.40%	74.18%	68.22%
<20 ms	83.80%	83.00%	80.37%	84.78%	83.21%
<30 ms	89.94%	90.37%	86.45%	88.59%	88.96%
<40 ms	93.85%	93.20%	92.06%	93.48%	93.18%
<50 ms	94.97%	96.88%	96.73%	96.47%	96.41%
>50 ms	5.03%	3.12%	3.27%	3.53%	3.59%

Table 8: *Segmentation accuracy on the COGEN database using the MBROLIGN and Hybrid HMM/ANN-based methods.*

Nb Instances	2215	13255	5605	13635	
MBROLIGN	VV	VC	CC	CV	Total
<10 ms	61.20%	69.70%	66.30%	67.20%	67.63%
<20 ms	78.8%	85.20%	81.20%	82.30%	83.00%
<30 ms	88.20%	90.80%	89.40%	90.10%	90.13%
<40 ms	92.80%	94.30%	93.20%	94.10%	93.94%
<50 ms	95.10%	96.50%	95.4%	96.20%	96.12%
>50 ms	4.9%	3.50%	4.6%	3.8%	3.88%
Hybrid					
<10 ms	63.66%	74.20%	67.62%	74.20%	72.46%
<20 ms	80.81%	91.63%	86.89%	86.32%	88.09%
<30 ms	88.94%	95.55%	92.86%	91.86%	93.24%
<40 ms	93.91%	97.32%	95.90%	95.56%	96.18%
<50 ms	95.94%	98.53%	97.77%	97.73%	97.93%
>50 ms	4.06%	1.47%	2.23%	2.27%	2.07%

Table 9: *Segmentation accuracy on the LATINO-40 database using the MBROLIGN and Hybrid HMM/ANN methods.*

References

- K. Lenzo and A.W. Black, *Diphone Collection and Synthesis*, Proceedings of the International Conference on Speech and Language Processing, 2000, Beijing, China.
- P. Horak, *Automatic Speech Segmentation Based on DTW with the Application of the Czech TTS System*, Improvements in Speech Synthesis, Ed. by E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckwale, John Wiley and Sons Ltd., 2001, pp 331-340.
- A. J. Hunt and A.W. Black, *Unit Selection in a Concatenative Speech Synthesis system using Large Speech Database*, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1996, pp 373-376.
- L. R. Bahl and S. Balakrishnan-Aiyer and J. Bellegarda and M. Franz and P. Gopalakrishnan and D. Nahamoo and M. Novak and M. Padmanabhan and M. Picheny and S. Roukos, *Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task*, Proceedings of the International Conference on Acoustics Speech and Signal Processing, 1995 ,pp 41-44.
- J. K. Baker, *The Dragon System – An Overview*, IEEE Transactions on Acoustics Speech and Signal Processing, 1975 , pp 24-29.
- Lawrence R. Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, PTR Prentice Hall, 1993.
- M. J. Russel and K. M. Ponting and S. M. Peeling and S. R. Browning and J. S. Bridle and R. K. Moore and I. Galiano and P. Howell, *The ARM Continuous Speech Recognition System*, Proceedings International Conference on Acoustics Speech and Signal Processing, 1990, pp 69-72.
- P. C. Woodland and C. J. Leggetter and J. J. Odell and V. Valtchev and S. Young, *The 1994 HTK Large Vocabulary Speech Recognition System*, Pro-

ceedings of the International Conference on Acoustics, Speech and Signal Processing, 1995, pp 73-76.

- F. Malfrère and T. Dutoit, *High-Quality Speech Synthesis for Phonetic Speech Segmentation*, Proceedings of the European Conference on Speech Communication and Technology, 1997 , pp 2631-2634.
- O. Deroo, F. Malfrère and T. Dutoit, *Comparison of two different alignment systems: speech synthesis vs. Hybrid HMM/ANN*, Proc. European Conference on Signal Processing (EUSIPCO'98), Rhodes, Grece, pp. 1161-1164.
- C. Traber, *SVOX : The Implementation of a Text-to-Speech System for German*, PhD Thesis, 1995 , ETH Zurich.
- P. Cosi and D. Falavigna and M. Olmologo , *A Preliminary Statistical Evaluation of Manual and Automatic Segmentation*, Proceedings of the European Conference on Speech Communication and Technology, 1991, pp 693-696.
- B. Brugnara and D. Falavigna and M. Omologo, *Automatic Segmentation and Labeling of Speech based on Hidden Markov Models*, Speech Communication , 1993, pp 357-370.
- H.C. Leung and V.W. Zue , *A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech*, Proc. International Conference on Acoustics, Speech and Signal Processing, 1984, pp 2.7.1-2.7.4.
- A. Ljolje and M.D. Riley, *Automatic Segmentation and Labeling of Speech*, Proc. International Conference on Acoustics, Speech and Signal Processing, 1991, pp 473-476.
- D. Talkin and C.W Wightman, *The Aligner : Text-to-Speech Alignment Using Markov Models and a Pronunciation Dictionary*, Proceedings of Second ESCA/IEEE Workshop on Speech Synthesis, 1996, pp 89-92.
- B. Van Coile and L. Van Tichelen and A. Vostermans and J.W. Wang and M.

- Staessen, *PROTRAN: A Prosody Transplantation Tool for Text-to-Speech Applications*, 1994, Proceedings of ICSLP'94.
- L. E. Baum, *An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes*, *Inequalities* 3, 1972, pp 1-8.
- F. Jelinek, *Continuous Speech Recognition by Statistical Methods*, Proceedings of the IEEE, 1976, pp 532-536.
- C. S. Myers and L. R. Rabiner, *A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition*, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1981.
- H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.
- A. J. Robinson and F. Fallside, *A Recurrent Error Propagation Network Speech Recognition System*, *Computer Speech and Language*, 1991, pp 257-286.
- A. J. Robinson, *An Application of Recurrent Nets to Phone Probability Estimation*, Proceedings of the IEEE Transactions on Neural Network, 1994, pp 298-305.
- S. Dupont and C. Ris and O. Deroo and V. Fontaine, *Context Independent and Context Dependent Hybrid HMM/ANN Systems for Vocabulary Independent Tasks*, Proceedings of the European Conference on Speech Communication and Technology, 1997, pp 1947-1950.
- H. Franco and M. Cohen and N. Morgan and D. Rumelhart and V. Abrash, *Context-Dependent Connectionist Probability Estimation in a Hybrid Hidden Markov Model-Neural Net Speech Recognition System*, *Computer Speech and Language*, 1994, pp 211-222.
- M. Hochberg and G. D. Cook and S. Renals and A. J. Robinson and R. S. Schechtman, *The 1994 ABBOT Hybrid Connectionist-HMM Large Vocab-*

ulary Recognition System, Spoken Language Systems Technology Workshop, 1995, pp 170-176.

T. Dutoit and V. Pagel and N. Pierret and F. Bataille and O. Van Der Vreken, *The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Free for Use for Non Commercial Purposes*, International Conference on Speech and Language Processing, 1996, pp 1393-1396.

B. H. Juang and L. R. Rabiner and J. G. Wilpon, *On the Use of Bandpass Filtering in Speech Recognition*, Proceedings of the International Conference on Acoustics Speech and Signal Processing, 1986, pp 765-768.

H. Hermansky, *Perceptual Linear Predictive Analysis of Speech*, Journal of The Acoustic Soc. Am., 1990.

J. Koehler and N. Morgan and H. Hermansky and H. G. Hirsch and G. Tong, INTEGRATING RASTA-PLP INTO SPEECH RECOGNITION, Proceedings of the International Conference on Acoustics Speech and Signal Processing, Adelaide, Australia, april 1994, pp I-421 - I-424

L. F. Lamel and J. L. Gauvain and M. Eskenazi, *BREF, a Large Vocabulary Spoken Corpus for French*, Proceedings of the European Conference on Speech Communication and Technology, 1991, pp 505-508.

D. B. Paul and J. Baker, The Design for the Wall Street Journal-based CSR Corpus", DARPA Speech and Language Workshop, 1992, Morgan Kaufmann Publishers.

G. Deville and O. Deroo and S. Gielen and H. Leich and J. Van Parys, *Automatic Detection and Correction of Pronunciation Errors for Foreign Language Learners : The DEMOSTHENES Application*, Proceedings of the European Conference on Speech Communication and Technology, 1999, pp 843-846.

- V. Zue and S. Seneff and J. Glass, *Speech Database Development : TIMIT and Beyond*, Speech Communication, 1990. pp 351-356.
- R. Carré and R. Descoudt and M. Eskénazi and J. Mariani and M. Rossi, *The French Language Database : Defining, Planning and Recording a Large Database*, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1984.
- L.Lee and R.Rose, *A Frequency Warping Approach to Speaker Normalization*, IEEE Transactions on Speech and Audio Processing, vol.6, No.1, 1998, pp. 49-60.