# Improvements on Automatic Speech Segmentation at the Phonetic Level

Jon A. Gómez and Marcos Calvo

Departament de Sistemes Informàtics i Computació,
Universitat Politècnica de València, Spain
{jon,mcalvo}@dsic.upv.es
http://elirf.dsic.upv.es/

**Abstract.** In this paper, we present some recent improvements in our automatic speech segmentation system, which only needs the speech signal and the phonetic sequence of each sentence of a corpus to be trained. It estimates a GMM by using all the sentences of the training subcorpus, where each Gaussian distribution represents an acoustic class, which probability densities are combined with a set of conditional probabilities in order to estimate the probability densities of the states of each phonetic unit. The initial values of the conditional probabilities are obtained by using a segmentation of each sentence assigning the same number of frames to each phonetic unit. A DTW algorithm fixes the phonetic boundaries using the known phonetic sequence. This DTW is a step inside an iterative process which aims to segment the corpus and re-estimate the conditional probabilities. The results presented here demonstrate that the system has a good capacity to learn how to identify the phonetic boundaries.

**Keywords:** automatic speech segmentation, phoneme boundaries detection, phoneme alignment

## 1 Introduction

The two main applications of speech segmentation at the phonetic level are text-to-speech synthesis and acoustic models training. For both purposes it is useful to have available as many labelled sentences as possible. Doing this labelling task by hand implies a great and very expensive effort. Additionally, as some authors point out, manual segmentations of a single corpus carried out by different experts can differ significantly, thus it is reasonable to use automatic segmentations in the previous applications. As an example, some researchers gave the same speech database to different human experts to segment it. Then, they evaluated the differences between the manual segmentations obtained. In [1], 97% of the boundaries within a tolerance interval of 20 ms were found, and 93% in [2].

There are some different approaches for performing automatic segmentation of speech corpora when the phonetic sequence of each sentence is available. Most of them are systems that operate in two stages: the first one is done by a phonetic recognizer based on Hidden Markov Models (HMM), which fixes the phonetic boundaries by using the Viterbi algorithm with forced alignment, and the second stage adjusts the phonetic boundaries. In [1, 3, 4] different pattern recognition approaches are proposed for local adjustment of boundaries. [5] presents an HMM-based approach where pronunciation variation rules are applied and a recognition network is generated for each sentence. Then a Viterbi search determines the most likely path and obtains an adapted phonetic transcription for each sentence. This process is repeated until the adapted phonetic transcriptions do not change any more. Initial phone HMMs are generated with flat-start training using the canonical transcriptions of the sentences.

A Dynamic Time Warping (DTW) based method that aligns the spoken utterance with a reference synthetic signal produced by waveform concatenation is proposed in [6]. The known phonetic sequence of each sentence is used to generate the synthetic signal. The alignment cost function depends on the pair of phonetic segment classes being aligned, and is computed taking a combination of acoustic features. In [7] a set of automatic segmentation machines are simultaneously applied to draw the final boundary time marks from the multiple segmentation results. Then, a candidate selector trained over a manually-segmented speech database is applied to identify the best time marks. In [8] several linear and nonlinear regression methods are used for combining multiple phonetic boundary predictions which are obtained through various segmentation engines.

An approach inspired in the minimum phone error training algorithm for automatic speech recognition [9] is presented in [10]. The objective of this approach is to minimize the expected boundary errors over a set of phonetic alignments represented as a phonetic lattice. A quite different approach, which is presented in [11], uses an extension of the Baum-Welch algorithm for training HMMs that use explicit phoneme segmentation to constrain the forward-backward lattice. This approach improves the accuracy of automatic phoneme segmentation and is even more computationally efficient than the original Baum-Welch.

A technique that modifies the topology of the HMMs in order to control the duration of the phonetic boundaries is presented in [12]. The prototype for all the phones is defined as a 5-state left-to-right topology with duration control states at each end. This topology improves the segmentation accuracy by reducing the probability of looping at the beginning and end states, as these model the boundaries between phonetic units. The acoustic vectors within the transition from one phonetic unit to the other are clustered at these states.

In this paper we present a technique for automatic speech segmentation at the phonetic level based on the same idea of altering the topology of the HMMs. Nevertheless, three differences should be noted: (a) we calculate the emission probabilities in a different way, (b) the forced alignment is performed by a DTW algorithm, and (c) we do not use manually segmented sentences for training. Emission probabilities are computed by combining acoustic probabilities with conditional probabilities estimated *ad hoc* [13, 14]. The conditional probabilities reflect the relation between the acoustic and the phonetic probability densities.

The estimation of these conditional probabilities is done by means of a progressive refinement iterative process which segments all sentences of the training set at every step. The initial values of the conditional probabilities are obtained by using a segmentation into equal parts, i.e., the segments assigned to each phonetic unit within a sentence are equally long. The acoustic probability densities are computed using a GMM (Gaussian Mixture Model), obtained as a result of a clustering process.

Next, we describe in Section 2 the recent improvements on the automatic speech segmentation system. Then, in Section 3, we show and comment the experimentation results. Finally, we conclude in Section 4.

## 2 System Improvements

The previous version of our system operated in three stages: (1) a coarse segmentation based on acoustic-phonetic rules was used to estimate the initial conditional probabilities, (2) the refinement of these conditional probabilities by means of an iterative procedure, and (3) a local adjustment of phonetic boundaries considering distinct criteria depending on the pair of consecutive phonetic units [14]. In this work, we present two improvements to this strategy. The first one consists in using HMMs with a little variation in the topology based on the idea presented in [12]. The topology is modified by having states without loops at each end to control the duration of the transitions between phonetic units. This improvement avoids the need for the coarse segmentation. The other improvement consists in the use of transitions between phonetic units as additional units.

The iterative procedure for progressive refinement is based on a DTW algorithm that automatically segments each sentence. This algorithm aligns the sequence of states with respect to the sequence of acoustic frames. The sequence of states of each sentence is obtained by concatenating the model of each phonetic unit according to the given phonetic sequence. There are two relevant features in the topology of the models: the total number of states and the number of duration control states. Figure 1 shows a model with 8 emitting states and 3 duration control states at both sides. It is important to highlight that each phonetic unit can have a different number of states according to its nature.

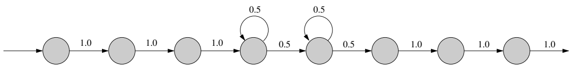

**Fig. 1.** *An 8 emitting states HMM with 3 duration control states at each side.*

Figure 2 shows the allowed movements inside the DTW matrix in an example of transition between two phonetic units, with one duration control state at each end. We can observe that horizontal movements are forbidden for duration control states, i.e., no loops are permitted. The diagonal movements are the only ones allowed for these states, as these movements represent the transition from one state to the next one. Vertical movements are always forbidden since it is inconsistent to assign one acoustic frame to more than one state.

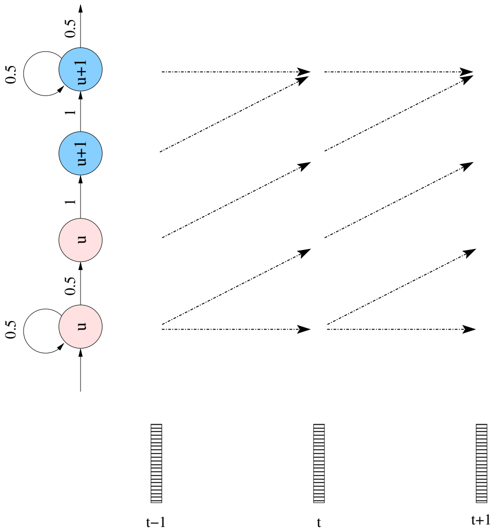

**Fig. 2.** *Example of possible movements in our DTW focused on the join between two phonetic units.*

The alignment cost function used in the DTW algorithm takes $p(x_t|e_i^u)$ as the emission probability, which represents the phonetic class-conditional probability density function of observing the acoustic frame $x_t$ given $e_i^u$, the $i$-th state of the phonetic unit $u$. This phonetic class-conditional probability density function is computed using the following formula

$$p(x_t|e_i^u) = \sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_i^u) \tag{1}$$

where $a$ is an acoustic class modelled by a Gaussian distribution, $A$ is the set of Gaussian distributions in the GMM which contains all the acoustic classes, $p(x_t|a)$ is the acoustic class-conditional probability density function of observing the acoustic frame $x_t$ given the acoustic class $a$, and $\Pr(a|e_i^u)$ is the conditional probability of the acoustic class $a$ given the state $e_i^u$ [13, 14]. The GMM is computed as the first step of the training process using all the acoustic frames of all the sentences of the training subcorpus. This acoustical clustering is performed by using the maximum likelihood estimation.

The initial values of the conditional probabilities are obtained from a segmentation of each sentence into equal parts. The progressive refinement stops when no variations are observed between the segmentations resulting from two consecutive iterations. As a further step, the transitions between each pair of phonetic units are added as new phonetic units, and new conditional probabilities are computed for the new set of units (original units plus transitions). The segmentation obtained in the last iteration of the previous progressive refinement process is used as the starting point for the estimation of the new set of conditional probabilities.

## 3 Experimentation

### 3.1 Speech Corpora

In order to carry out experiments for both Spanish and English, we chose two speech databases: *Albayzin* [15] and TIMIT [16]. The phonetic corpus from the *Albayzin* database was used for the Spanish experiments: 6,800 utterances (around six hours of speech) which we split into 1,200 sentences manually segmented and labelled that were used for testing and the remaining 5,600 sentences for training. No speakers appear in both subsets. The TIMIT database was used for the English experiments, which contains 6,300 utterances (approximately five hours of speech). In this case we used the suggested training/test subdivision.

The same acoustic parameters were used on both databases. Each acoustic frame was formed by a 39-dimensional vector composed by the normalized energy, the first 12 Mel frequency cepstral coefficients, and their first and second time derivatives. Each acoustic frame was obtained using a 20 ms Hamming window every 5 ms.

### 3.2 Evaluation Criteria

The most widely used evaluation criterion to measure the accuracy of an automatic segmentation with respect to a manual one is the percentage of boundaries which error is within a tolerance. Usually, it is calculated for a range of tolerances [1, 2, 8].

As discussed in the introduction, some researchers have wondered if a manual segmentation could be a valid reference [1, 2]. To evaluate this, they gave the same speech corpus to different human experts asking them to annotate it, and then evaluated the differences among the manual segmentations. In the study presented in [1], 97% of the boundaries were found within a tolerance of 20 ms and in [2] 93%. Thus, we can interpret these results as an upper bound for the accuracy of automatic segmentations, since a system that reaches 100% compared with a manual segmentation will at least differ around 5% from another manual segmentation for the same speech database.

### 3.3 Experimental Results

Our system has been evaluated using different combinations of the number of emitting states ($E$) and duration control states ($B$). Table 1 presents the results obtained using different $E \times B$ topologies. Results show that the use of duration control states lead to a significant improvement when the tolerance ranges from 5 to 20 ms. This improvement is bigger when the tolerance interval is more restrictive. For example, using the *Albayzin* corpus, if $E = 7$ then the segmentation accuracy improves from 58.5% to 67.8% for a tolerance error of 10 ms as $B$ increases, and from 85.2% to 89.1% for 20 ms.

**Table 1.** *Percentage of correctly fixed phonetic boundaries for a range of tolerances.*

| | *Albayzin* | | | | | | TIMIT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Topology | Tolerance in ms | | | | | | Tolerance in ms | | | | | |
| $E \times B$ | 5 | 10 | 15 | 20 | 30 | 50 | 5 | 10 | 15 | 20 | 30 | 50 |
| 5x1 | 33.0 | 58.5 | 74.9 | 85.3 | 94.6 | 98.7 | 25.5 | 46.6 | 62.0 | 72.7 | 88.0 | 97.7 |
| 5x2 | 36.8 | 62.6 | 78.6 | 87.5 | 94.7 | 98.5 | 22.4 | 43.7 | 61.9 | 74.8 | 89.6 | 97.8 |
| 6x2 | 37.1 | 64.4 | 80.0 | 87.9 | 95.2 | 98.8 | **29.5** | **53.6** | 69.9 | 80.3 | 91.6 | 97.9 |
| 7x0 | 31.7 | 58.5 | 75.5 | 85.2 | 94.1 | 98.3 | 24.4 | 44.9 | 60.8 | 72.3 | 88.0 | 97.9 |
| 7x1 | 33.6 | 61.0 | 77.3 | 85.8 | 94.4 | 98.5 | 24.4 | 45.2 | 62.3 | 74.3 | 89.5 | 98.1 |
| 7x2 | 36.2 | 63.0 | 78.6 | 86.9 | 95.1 | 98.7 | 28.5 | 52.1 | 68.9 | 79.8 | 91.8 | 98.2 |
| 7x3 | 40.9 | 67.8 | 82.1 | 89.1 | 95.6 | 98.9 | 24.7 | 47.8 | 66.6 | 78.6 | 91.2 | 98.1 |
| 8x3 | 40.5 | 67.5 | 82.1 | **89.5** | **96.2** | **99.2** | 27.8 | 51.9 | 70.7 | **82.7** | 93.6 | 98.5 |
| 9x2 | 39.8 | 66.8 | 81.1 | 88.5 | 95.7 | 98.9 | 28.6 | 52.2 | 69.0 | 79.8 | 91.6 | 97.7 |
| 9x3 | 38.1 | 66.0 | 81.5 | 89.0 | 96.1 | 99.2 | 28.2 | 52.0 | **70.8** | 82.6 | **93.8** | **98.6** |
| 9x4 | **44.0** | **70.3** | **82.8** | 89.4 | 95.8 | 99.0 | 25.4 | 49.9 | 69.3 | 81.5 | 92.7 | 98.2 |
| 10x4 | 42.5 | 68.9 | 82.2 | 88.9 | 95.8 | 99.0 | 26.3 | 50.1 | 68.2 | 79.9 | 91.6 | 98.1 |

As mentioned above, our system does not use any manual segmentation for bootstraping. Starting from a blind segmentation of the sentences into equal parts, the learning process converges in less than 20 iterations for all the topologies considered.

We used a subsampling rate of 200 Hz, so, an HMM with 8 emitting states implies a minimum duration for each phonetic unit of 40 ms, which is longer than usual for some of them. Thus, different topologies were used for voiced plosives /b/, /d/ and /g/ when the topology of the remaining phonetic units is larger than 5 states. In the experiments performed with the *Albayzin* corpus, a $5 \times 2$ topology was used for these units. The results improved significantly thanks to this shorter topology. The structure of voiceless plosives /p/, /t/ and /k/ was not different from the topologies used for the rest of units, since their preceding silence is properly clustered by the HMM states. Silences were considered a special case and were always modelled with a $3 \times 0$ topology.

Since in the TIMIT corpus the voiceless plosives are preceded by a unit representing the closure, a shorter topology was needed for these units. A $3 \times 1$ topology was used for /b/, /d/, /g/, /p/, /t/, and /k/.

Additionally, we also considered adding the transitions between pairs of consecutive phonetic units as extra ones. Table 2 shows the results obtained when a $6 \times 2$ topology was used for all units except plosives, which were modelled with 4 × 1 for *Albayzin* and 3 × 1 for TIMIT. The silences were modelled with a 3 × 0 topology for both corpora. In the case of the *Albayzin* corpus no significant improvements are observed. However, experiments with the TIMIT corpus show small improvements for tolerances of 5 and 10 ms. Also, a significant improvement can be observed when using the manually segmented sentences of the training subcorpus to initialize of the conditional probabilities.

**Table 2.** *Percentage of correctly fixed phonetic boundaries when transitions were used. For the* TIMIT *corpus results when using the manual segmentation for training are also presented. No manual segmentation for training is available in the Albayzin corpus.*

| | *Albayzin* | | | | | | TIMIT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Using | Tolerance in ms | | | | | | Tolerance in ms | | | | | |
| manual | 5 | 10 | 15 | 20 | 30 | 50 | 5 | 10 | 15 | 20 | 30 | 50 |
| No | 40.6 | 68.7 | 83.2 | 90.5 | 96.4 | 99.3 | 31.5 | 55.8 | 71.0 | 81.1 | 92.3 | 98.2 |
| Yes | | | | | | | 44.1 | 70.3 | 81.9 | 88.2 | 94.8 | 98.7 |

## 4 Conclusions

We have presented here an automatic segmentation technique that combines three ideas. The first one consists in using duration control states at each end of every HMM as well as increasing the number of emitting states. The second, detailed in Section 2, deals with the way emission probabilities are calculated. The third idea consists in using a DTW algorithm to align the sequence of states against the sequence of acoustic frames.

The goal of our approach is to automatically segment speech corpora that can be useful to train acoustic models without the need for manually segmented and labelled sentences. The obtained segmentation accuracy for the *Albayzin* corpus in both kinds of experiments is around 90% within a tolerance of 20 ms. This enables our system to be used for the planned purposes, namely, acoustic models training and concatenative text-to-speech synthesis.

The results achieved with the TIMIT corpus without using the manually segmented sentences for training are similar to the ones obtained by other researchers referenced above using standard HMM and the manually segmented sentences. We have also used the transitions between phonetic units, but this only improves the segmentation accuracy for tolerances of 5 and 10 ms. When our system is trained using the manually segmented sentences the results are even better.

## Acknowledgments

This work was supported by the Spanish MICINN under contract TIN2008-06856-C05-02.

## References

1. Toledano, D. T., Hernández Gómez, L. and Villarrubia Grande, L.: Automatic Phonetic Segmentation. In: IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 617–625, November 2003.
2. Kipp, A., Wesenick, M.B. and Schiel F.: Pronunciation modelling applied to automatic segmentation of spontaneous speech. In: Proceedings of Eurospeech, 1997, pp. 2013–1026, Rhodes, Greece.
3. Sethy, A., Narayanan, S.: Refined Speech Segmentation for Concatenative Speech Synthesis. In: Proceedings of ICSLP, 2002, pp. 149–152, Denver, Colorado, USA.
4. Jarify, S., Pastor, D., Rosec, O.: Cooperation between global and local methods for the automatic segmentation of speech synthesis corpora. In: Proceedings of Interspeech, 2006, pp. 1666–1669, Pittsburgh, Pennsylvania, USA.
5. Romsdorfer, H., Pfister, B.: Phonetic Labeling and Segmentation of Mixed-Lingual Prosody Databases. In: Proceedings of Interspeech, 2005, pp. 3281–3284, Lisbon, Portual.
6. Paulo, S., Oliveira, L.C.: DTW-based Phonetic Alignment Using Multiple Acoustic Features. In: Proceedings of Eurospeech, 2003, pp. 309–312, Geneva, Switzerland.
7. Park, S.S., Shin, J.W., Kim, N.S.: Automatic Speech Segmentation with Multiple Statistical Models. In: Proceedings of Interspeech, 2006, pp. 2066–2069, Pittsburgh, Pennsylvania, USA.
8. Mporas, I. and Ganchev, T. and Fakotakis, N.: Speech segmentation using regression fusion of boundary predictions. Computer Speech and Language, Academic Press Ltd., London, UK, vol. 24, pp. 273–288, April, 2010.
9. Povey, D., Woodland, P.C.: Minimum Phone Error and I-smoothing for improved discriminative training. In: Proceedings of ICASSP, 2002, pp. 105–108, Orlando, Florida, USA.
10. Kuo, J.W., Wang, H.M.: Minimum Boundary Error Training for Automatic Phonetic Segmentation. In: Proceedings of Interspeech, 2006, pp. 1217–1220, Pittsburgh, Pennsylvania, USA.
11. Huggins-Daines, D., Rudnicky, A.I.: A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training. In: Proceedings of Interspeech, 2006, pp. 1205–1208, Pittsburgh, Pennsylvania, USA.
12. Ogbureke, Kalu U., Carson-Berndsen, Julie: Improving initial boundary estimation for HMM-based automatic phonetic segmentation. In: Proceedings of Interspeech, 2009, pp. 884–887, Brighton, U.K.
13. Gómez, J.A. and Castro, M.J.: Automatic Segmentation of Speech at the Phonetic Level, In: Structural, Syntactic, and Statistical Pattern Recognition, volume 2396 of *LNCS*, pp. 672–680. Springer-Verlag, 2002.
14. Gómez, Jon A. and Sanchis, Emilio and Castro-Bleda, María J.: Automatic speech segmentation based on acoustical clustering. In: Structural, Syntactic, and Statistical Pattern Recognition, LNCS, vol. 6218, pp. 540–548, Springer-Verlag, Berlin, Heidelberg (2010).
15. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Nadeu, C.: Albayzin Speech Database: Design of the Phonetic Corpus. In: Proceedings of Eurospeech, 1993, volume 1, pages 653–656. Berlin (Germany), September 1993.
16. TIMIT Acoustic-Phonetic Continuous Speech Corpus, National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-5050651996, October 1990.