FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Automatic Phonetic Segmentation

Jon Ander Gómez Adrián

Departmento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
jon@dsic.upv.es

Abstract

This paper presents an approach to automatic segmentation of speech corpora. The availability of sufficiently precise labelled sentences can avoid the need for a segmentation by human experts. The goal of this process is to prepare speech corpora both for training acoustic models and for concatenative text to speech synthesis.

Our system only needs the speech signal and the phonetic sequence for each sentence of a corpus. It estimates a GMM by using all sentences, where each Gaussian distribution represents an acoustic class. Then it combines the probability densities of each acoustic class with a set of conditional probabilities in order to estimate the probability densities of the states of 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. A flat start setup is used to give initial values to the conditional probabilities.

Index Terms: automatic speech segmentation, phoneme boundaries detection, phoneme alignment

1. Introduction

The two main applications of phonetic level segmentation are text to speech synthesis and acoustic models training. In both cases it is useful to have as many labelled sentences as possible. Doing this labelling task by hand implies a great effort that can be very expensive. Furthermore, as some authors point, manual segmentations of a single corpus carried out by different experts can have significant differences, thus it is reasonable to use automatic segmentations. As an example, some researchers have given 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 several different approaches to the automatic segmentation of sentences when the phonetic sequence is available. Most of them are systems in two stages: the first one is performed by a Hidden Markov Model (HMM) based phonetic recognizer using the Viterbi forced alignment, and the second one adjusts the phonetic boundaries. In [1, 3, 4] different pattern recognition approaches are proposed for the local adjustment of boundaries. [5] presents a 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 which 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 is computed using a combination of acoustic features depending on the pair of phonetic segment classes being aligned. 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.

An approach inspired by the minimum phone error training algorithm for automatic speech recognition [8] is presented in [9]. 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 is presented in [10], which uses an extension of the Baum-Welch algorithm for training HMM that uses explicit phoneme segmentation to constrain the forward and backward lattice. This approach improves the accuracy of automatic phoneme segmentation and is even more computationally efficient than the original Baum-Welch.

A technique which modifies the topology of the HMM to control for duration is presented in [11]. The prototype for all phones is defined as a 5-state left-right topology with duration control states at each end. This topology improves segmentation accuracy by reducing the probability of remaining in the beginning and end states as these states model the boundaries between phonetic units. The acoustic vectors at the transition from one phonetic unit to the other are clustered at these states.

In this paper we present a phonetic level automatic speech segmentation technique based on the same idea of altering the topology of the HMM. Nevertheless, we calculate the emission probabilities in a different way, the forced alignment is performed by means of a DTW algorithm and we do not use any manually segmented sentences. Emission probabilities are computed by combining acoustic probabilities with conditional probabilities estimated *ad hoc*. 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 an iterative process of progressive refinement which segments all sentences of the training set at every step. The initial values given to the conditional probabilities are calculated using a flat start setup, and the acoustic probability densities are computed from a GMM (Gaussian Mixture Model) obtained as a result of a clustering process.

Next, we describe in Section 2 the proposed approach for automatic speech segmentation. Then, in Section 3, we show and comment the experimentation results. Finally, we conclude in Section 4.

2. System description

A DTW algorithm to automatically segment each sentence is used. This algorithm aligns the sequence of states with respect to the sequence of acoustic frames. The sequence of states is composed by concatenating the model of each phonetic unit from the known phonetic sequence. There are two relevant constraints on the topology of the models, that are the 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, similar to the ones proposed in [11]. The number of states sets the minimum number of frames assigned to each phonetic unit.

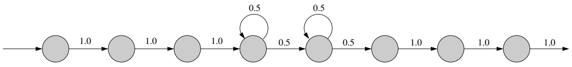

Figure 1: *An 8 emitting states HMM with 3 duration control states at each side.*

The alignment cost function used in the DTW algorithm uses $\Pr(e_i^u|x_t)$ as the emission probability, which represents the posterior probability of each state given an acoustic vector, where e_i^u is the i-th state of the phonetic unit u. For each acoustic frame x_t we obtain another vector with the phonetic level probabilities $\{\Pr(e_i^u|x_t)\}\ \forall u \in U,\ i = 1..E(u)$, where U is the set of phonetic units and $E(u)$ is the number of states of the phonetic unit u. Applying this process to each frame of an utterance we obtain as a result a sequence of vectors with the probability of each state of each phonetic unit.

2.1. Acoustic probabilities

The acoustic probabilities are computed from a GMM where each Gaussian distribution represents an acoustic class. The GMM is estimated by means of a clustering procedure using as training samples all the acoustic vectors from every training sentence. The unsupervised learning of the means and diagonal covariances for each acoustic class has been done by maximum likelihood estimation as described in [12].

The underlying idea of our approach is based on the fact that, once we transform the waveform into frames (d-dimensional acoustic vectors), they are distributed into a region of $\mathbb{R}^d$, in such a way that more dense subregions are formed according to similar acoustic phonetic features. The dense subregions can be related to different acoustical manifestations of speech. Each phonetic unit can have many acoustically different ways of being pronounced due to, among other phenomena, the mood and the accent of the speaker. Context is also an important factor that affects to the way a phoneme is pronounced, previous and following phonemes influence the one being uttered. In addition, not all the possible acoustic manifestations are related to an only phoneme, but many of them fall in the intersection of two or more phonemes. So, we can conclude that the subregions of each phoneme are neither isolated nor continuous.

In short, we can consider that the phonetic units are distributed in overlapped subregions inside $\mathbb{R}^d$, and that the natural acoustic classes allow us to model more precisely the region of this space where acoustic frames are distributed. It is easy to see that the number of acoustic classes will be much larger than the number of phonetic units.

2.2. Phonetic probabilities

To take into account the different degrees of relation between a phonetic class and a phonetic unit we have used conditional probabilities estimated for this goal. So, for each state we have $\Pr(a|e_i^u)$, which represents the conditional probability that the acoustic class a has occurred having that the phonetic unit u has been pronounced and its internal state e_i^u is active.

We can compute the class-conditional probability density function of observing the acoustic frame x_t assuming that the acoustic class a has been manifested, $p(x_t|a)$, according to the GMM. Nevertheless, we need the phonetic-conditional probability density function of observing the acoustic frame x_t given the state e_i^u, that can be calculated as follows:

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

Applying the Bayes rule we can obtain the posterior probability of each phonetic state given a frame:

$$\Pr(e_i^u|x_t) = \frac{p(x_t|e_i^u)\pi(e_i^u)}{\sum\limits_{v \in U} \sum\limits_{j=1}^{E(v)} p(x_t|e_j^v)\pi(e_j^v)} \tag{2}$$

where $\pi(\cdot)$ is the prior probability of each state of each phonetic unit. In our approach we consider that all the prior probabilities are the same, so we can eliminate them. Taking this into account and expanding Equation 2 according to Equation 1 we have:

$$\Pr(e_i^u|x_t) = \frac{\sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_i^u)}{\sum\limits_{v \in U} \sum\limits_{j=1}^{E(v)} \left(\sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_j^v) \right)} \tag{3}$$

The conditional probabilities $\Pr(a|e_i^u)$ can be estimated either in a supervised way, using a manually segmented and labelled corpus, or in an unsupervised way, using an iterative process of progressive refinement like the one proposed here.

The initial values of the conditional probabilities are calculated using a flat start setup. Then the iterative process that re-estimates the conditional probabilities starts and goes on until there are no significant changes.

2.3. DTW state alignment

The DTW algorithm used to align the sequence of states against the sequence of acoustic frames obtains the phonetic level segmentation. For each sentence we can build the sequence of states by concatenating the models of the phonetic units that were pronounced. It is important to highlight that each phonetic unit can have a different number of states according to its nature. Figure 2 shows the allowed movements inside the DTW matrix for an example corresponding to the join between two phonetic unit models, with one duration control state at each end. We can observe that horizontal movements are forbidden for the duration control states, which only allow diagonal movements. Vertical movements are not allowed, since it would imply that an only frame is assigned to more than one state.

3. Experimentation results

In this section we describe the performed experiments and the obtained results. First, we present the speech corpus used for testing our system and comment the evaluation criteria. Next, the results for different combinations of the total number of

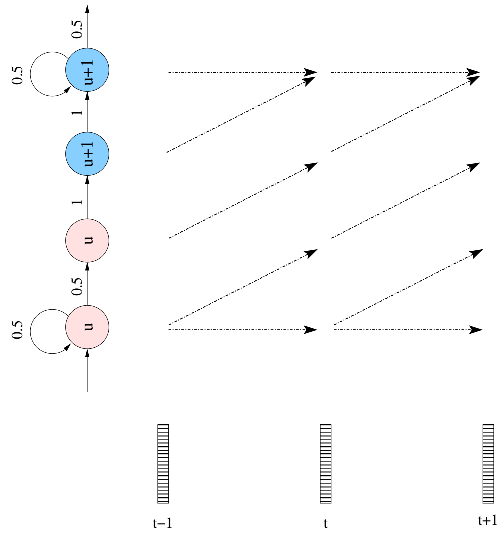

Figure 2: *Example of possible movements in our DTW focused on the join between two phonetic units.*

states and the number of duration control states at each end are presented. Our goal was to find the optimal topology for all phonetic units, so we repeated the training and testing processes using different configurations. When the best configurations were detected, we focused on those phonetic units we considered should have a different number of states, so we tried to achieve better results modifying their particular topology.

We also tested what happened when we did not allow the re-estimation of the transition probabilities of central states. The obtained results show that re-estimating the transition probabilities runs worse than not re-estimating them for every topology tested. Only results obtained with the best modality are presented here.

3.1. Speech corpus

The phonetic subcorpus from *Albayzin* database [14] that was used for the experiments is composed by 6,800 utterances (around six hours of speech) obtained by making groups from a set of 700 different sentences uttered by 40 different speakers. 1,200 sentences manually segmented and labelled were used for testing and the remaining 5,600 sentences were used for training. There are no common speakers between the training and test subcorpora.

Each acoustic frame is 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.

It is worth to say that we did not use the original training and test subsets that had the database because all the manually segmented and labelled sentences were included in the training subcorpus. So, we used the subset of 1,200 manually segmented and labelled sentences for testing and the 5,600 remaining sentences for training.

3.2. Evaluation criteria

The evaluation criteria most widely used in the literature is to measure the agreement of the obtained segmentation with respect to a manual segmentation. Usually the percentage of boundaries whose error is within a tolerance is calculated for a range of tolerances [1, 2, 13].

As discussed in the introduction, some researchers have wondered whether or not a manual segmentation is a valid reference [1, 2]. To evaluate it, they gave the same speech database to different human experts to segment it, and then evaluated the differences between them. In the study presented in [1], 97% of the boundaries within a tolerance of 20 ms were found and in [2] 93%. We interpret this agreement as the maximum accuracy for a segmentation system, since a system that reaches 100% compared with a manual segmentation will at least differ around 95% with another manual segmentation for the same speech database.

3.3. Experimental results

Our system has been evaluated for different combinations of the number of emitting states and duration control states. Table 1 presents the results obtained using different $E \times B$ topologies, where E represents the number of emitting states and B the number of duration control ones. Furthermore, Figure 3 shows a graphic representation of the same results, where a significant improvement is easily observed when tolerance increases from 10 to 20 ms.

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

	Tolerance en ms					
Topology	5	10	15	20	30	50
3x0	23.8	47.2	66.9	80.6	92.0	97.9
3x1	26.6	51.9	70.0	82.0	92.4	97.8
5x0	28.4	52.5	71.3	83.3	93.4	98.3
5x1	32.3	58.6	76.5	87.0	94.7	98.6
5x2	36.7	62.6	78.7	88.0	94.8	98.5
6x2	37.0	63.1	79.2	88.5	95.2	98.8
7x0	32.6	58.8	75.9	85.9	94.9	98.6
7x1	34.0	61.3	79.0	88.5	95.5	98.8
7x2	34.6	62.1	79.5	88.7	95.4	98.7
7x3	35.7	63.6	80.5	89.1	95.4	98.8
8x3	40.3	67.2	81.9	89.0	95.8	99.0
9x2	37.4	65.7	81.2	88.5	92.5	95.3
9x3	39.4	67.6	82.2	89.0	95.7	98.9
9x4	42.4	69.2	82.0	88.6	95.3	98.8

Results show a significant improvement when duration control states at each end are used. Also we can observe that the more restrictive a tolerance interval is, the more relevant is the improvement we achieve. For example, if $E = 7$ then the segmentation accuracy improves from 58.8% to 63.6% as B increases, for a tolerance error of 10 ms, and from 85.9% to 89.1% for 20 ms. By observing the results for different values of E we can detect a better performance when all the states except the central ones are duration control states.

As mentioned above, our system begins the learning process from a flat start setup and then iterates to re-estimate the conditional probabilities which relate the acoustic probability densities to the phonetic ones. Figure 4 shows the evolution of segmentation accuracy for several topologies within a tolerance interval of 20 ms. No significant improvements are obtained from 15th step, and we can clearly see the difference of segmentation accuracy when the 7×0 topology was used.

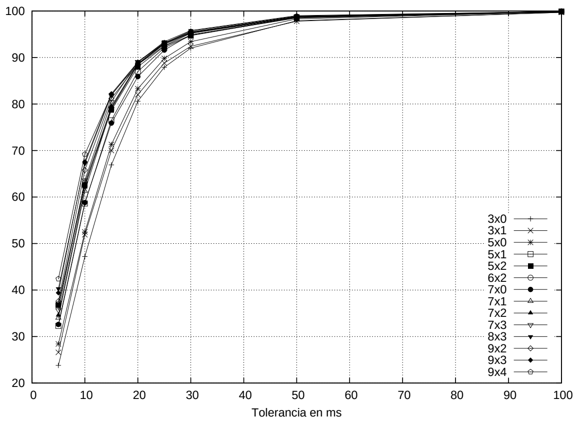

Figure 3: *Evolution of segmentation accuracy in function of tolerance error.*

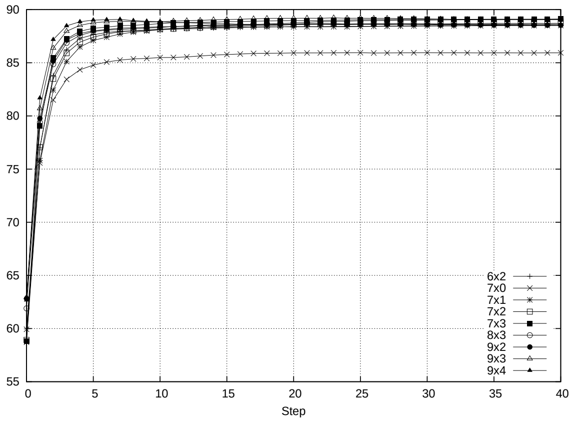

Figure 4: *Evolution of segmentation accuracy in function of iterative steps with a tolerance error of 20 ms.*

Taking into account that the subsampling rate is 200 Hz, a HMM with 8 emitting states forces a minimum phone duration of 40 ms, which is longer than usual for some phonetic units. The topologies of voiced plosives /b/, /d/ and /g/ differs from the topologies of the remaining units when using models with more than 7 emitting states. In this particular case, a 5 × 2 topology was used and the results improved significantly when this change was applied. The topologies of voiceless plosives /p/, /t/ and /k/ were not different from the topologies used for the rest of units. The burst of these plosives is always preceded by a short silence. So, voiceless plosives do not need a special topology because the frames of preceding silence are properly clustered by the HMM states. Finally, the phonetic unit representing silences is considered a special case, for which we used a 3 × 0 topology.

4. Conclusions

We have presented here an automatic segmentation technique which combines three ideas. The first one consists in using duration control states at each end of each HMM and in increasing the number of emitting states. This idea improves significantly the segmentation accuracy as it was shown by some researchers [11]. The second one, 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 main goal of our approach is to automatically segment speech corpora for training acoustic models without making use of any subset of manually segmented and labelled sentences. A segmentation accuracy close to 90% within a tolerance of 20 ms enables our system to be used for this purpose. In addition, our system can be useful for concatenative text-to-speech synthesis.

5. Acknowledgments

I am very grateful to Marcos Calvo Lance, he supported me correcting English text.

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

6. 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, pages 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, pages 2013–1026, Rhodes, Greece.

[3] Sethy, A., Narayanan, S., "Refined Speech Segmentation for Concatenative Speech Synthesis", in Proceedings of ICSLP, 2002, pages 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, pages 1666–1669, Pittsburgh, Pennsylvania, USA.

[5] Romsdorfer, H., Pfister, B., "Phonetic Labeling and Segmentation of Mixed-Lingual Prosody Databases", in Proceedings of Interspeech, 2005, pages 3281–3284, Lisbon, Portual.

[6] Paulo, S., Oliveira, L.C., "DTW-based Phonetic Alignment Using Multiple Acoustic Features", in Proceedings of Eurospeech, 2003, pages 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, pages 2066–2069, Pittsburgh, Pennsylvania, USA.

[8] Povey, D., Woodland, P.C., "Minimum Phone Error and I-smoothing for improved discriminative training", in Proceedings of ICASSP, 2002, pages 105–108, Orlando, Florida, USA.

[9] Kuo, J.W., Wang, H.M., "Minimum Boundary Error Training for Automatic Phonetic Segmentation", in Proceedings of Interspeech, 2006, pages 1217–1220, Pittsburgh, Pennsylvania, USA.

[10] Huggins-Daines, D., Rudnicky, A.I., "A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training", in Proceedings of Interspeech, 2006, pages 1205–1208, Pittsburgh, Pennsylvania, USA.

[11] Ogbureke, Kalu U., Carson-Berndsen, Julie, "Improving initial boundary estimation for HMM-based automatic phonetic segmentation", in Proceedings of Interspeech, 2009, pages 884–887, Brighton, U.K.

[12] Duda, R. O., Hart, P. E. and Stork, D. G., *"Pattern Classification"*, John Wiley and Sons, second edition, 2001.

[13] Mporas, I., Ganchev, T., and Fakotakis, N., "A Hybrid Architecture for Automatic Segmentation of Speech Waveforms", IEEE ICASSP 2008, pages 4457–4460, Las Vegas, USA.

[14] 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.