Automatic Phonetic Alignment and Its Confidence Measures

Sérgio Paulo and Luís C. Oliveira

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

Abstract. In this paper we propose the use of an HMM-based phonetic aligner together with a speech-synthesis-based one to improve the accuracy of the global alignment system. We also present a phone duration-independent measure to evaluate the accuracy of the automatic annotation tools. In the second part of the paper we propose and evaluate some new confidence measures for phonetic annotation.

1 Introduction

The flourishing number of spoken language repositories has pushed speech research in multiple ways. Much of the best speech recognition systems rely on models created with very large speech databases. Research into natural prosody generation for speech synthesis is, nowadays, another important issue that uses large amounts of speech data. These repositories have allowed the development of many corpus-based speech synthesizers in the recent years, but they need to be phonetically annotated with a high level of precision. However, manual phonetic annotation is a very time-consuming task and several approaches have been taken to automate this process. Although state-of-the-art segmentation tools can achieve very accurate results, there are always some uncommon acoustic realizations or some kind of noise that can badly damage the segmentation performance for a particular file. With the increasing size of speech databases manual verification of every utterance is becoming unfeasible, thus, some confidence scores must be computed to detect possible bad segmentations within each utterance. The goal of this work is the development of a robust phonetic annotation system, with the best possible accuracy, and the development and evaluation of confidence measures for phonetic annotation process. This paper is divided into 4 sections, the section 2 describes the development of the proposed phonetic aligner. In the following section (section 3), we describe and evaluate the proposed confidence measures, and the conclusions in the last section.

J. L. Vicedo et al. (Eds.): EsTAL 2004, LNAI 3230, pp. 36–44, 2004.

2 Automatic Segmentation Approaches

Automatic phonetic annotation is composed of two major steps, the determination of the utterance phone sequence, the sequence produced by the speaker during the recording procedure, and the temporal location of the segment boundaries (phonetic alignment). Several phonetic alignment methods have been proposed, but the most widely explored techniques are based either on Hidden Markov Models (HMM) used in forced alignment mode [1] or on dynamic time alignment with synthesized speech [2]. The main reason of the superiority of two techniques is their robustness and accuracy, respectively. An HMM-based aligner consists of a finite state machine that has a set of state occupancy probabilities in each time instant and a set of inter-state transition probabilities. These probabilities are computed using some manually or automatically segmented data (training data). On the other hand, the speech-synthesis-based aligners are based on a technique used in the early days of the speech recognition. A synthetic speech signal is generated with the expected phonetic sequence, together with the segment boundaries. Then, some spectral features are computed from the recorded and the generated speech signals, and finally the Dynamic Time Warping (DTW) algorithm [3] is applied to compute the aligned path between the signals for which there is a better match between the spectral features. The reference signal segment boundaries are mapped into the recorded signal using this alignment path. A comparison between the results of HMM-based and speech-synthesis-based segmentation [4] has showed that in general (about 70% of times) the speech-synthesis-based segmentation is more accurate than the HMM-based one, however, it tends to generate few large boundary errors (when it fails it fails badly). This means that the HMM-based phonetic aligners are more reliable.

The lack of robustness of the speech-synthesis-based aligners as well as its better boundary location accuracy suggested the development of an hybrid system, a system as accurate as the speech-synthesis-based aligner and with the robustness of the hmm-based aligners.

2.1 Speech Synthesis Based Phonetic Aligners

The first conclusion taken from the usage of some commonly used speech-synthesis-based aligners is that the acoustic features does not prove to be equally good for locating the boundaries for every kind of phonetic segment. For instance, although the energy is, in general, a good feature to locate the boundary between a vowel and a stop consonant, it performed poorly on locating the boundary between two vowels. Thus, some experiments were performed with multiple acoustic features and multiple segment transitions to find the best acoustic features to locate the boundaries between each different pair of phonetic segments. This acoustic feature selection considerably increased the robustness of the resulting aligner. The reference speech signal was generated using the Festival Speech Synthesis System [5] using a Portuguese voice recorded at our lab. A detailed description of this work can be found in [6].

2.2 HMM Based Phonetic Aligners

Once the speech-synthesis-based aligner was built with a good enough robustness, it was used to generate the training data for the HMM-based aligner. Given the amount of available training data, context-independent models were chosen for the task. Figure 1 shows the different phone topologies. The upper one is used for all phonetic segment but the silence, semi-vowels and shwa. The central topology is used to represent segments with short durations like the semi-vowels and shwa, by allowing a skip between and first and last states. The silence model is the lower one. In this case a transition from the first state to the last as well as another one from the last to the first state can be observed, this can be used to model very large variations on the duration of the silences in the speech database. Each model states consists of a set of eight gaussian mixtures. The adopted features were the Mel-Frequency Cepstrum Coefficients, their first and second order differences and the energy and its first and second differences. Each frame is spaced by 5-miliseconds , with a 20-milisecond long window. The training of the model was preformed by using the HTK toolkit.

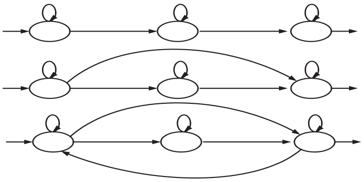

Fig. 1. Three HMM topologies were used for the different kinds of phonetic segments. The upper one is the general model, the central topology in used for semi-vowels and shwa, and the last one for the silence

2.3 Segment Boundary Refinement

As expected, using the HMM-based aligner, a more robust segmentation was obtained. The next step was to use our speech-synthesis-based aligner to refine the location of the segment boundaries.

2.4 Segmentation Results

Two of the most common approaches to evaluate the segmentations' accuracy is to compute of the phonetic segment percentage that have boundary location errors less than a given tolerance (often 20 ms), or the root mean square error of the boundary locations. Although these can be good predictors for aligners' accuracy, it is clear that an error of about 20 ms in a 25-ms long segment is much more serious that the same error in a 150-ms long segment. In the first case the segment frames are almost always badly assigned. This way, a phone-based duration-independent measure is proposed to evaluate the aligners' accuracy, that is to determine the percentage of well assigned frames, within the segment. We will call it the *Overlap Rate* (*OvR*). Fig. 2 illustrates the computation of this measure. Given a segment, a reference segmentation (RefSeg), and the segmentation to be evaluated(AutoSeg), OvR is the ratio between the number of frames that belong to that segment in both the segmentations ($Common\_Dur$ in the fig. 2) and the number of frames that belong to the segment in one segmentation, at least(Dur_max if the fig. 2). The following equation illustrates the computation of OvR:

$$OvR = \frac{Common\_Dur}{Dur\_max} = \frac{Common\_Dur}{Dur\_ref + Dur\_auto - Common\_Dur} \qquad (1)$$

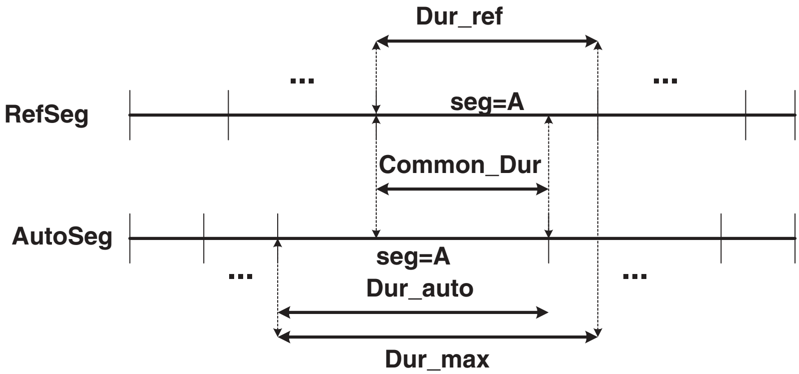

Fig. 2. Graphical representation of the quantities involved in the computation of the Overlap Rate

Regarding the equation 1, one can realize that if, for example, a phone duration in the reference segmentation differs considerably from its duration in the other segmentation, the OvR quantity takes a very small value. Let X be the Dur_ref, Y the Dur_auto and z the $Common\_Dur$ of Fig. 2, and suppose $X \leq Y$, thus:

$$0 \leq OvR = \frac{z}{X + Y - z} \leq \frac{X}{Y} \qquad (2)$$

since the number of common frames (z) is at most the same as the minimum number of frames in the two annotations of the given segment. This way, one can conclude that this measure is duration independent, and is able to produce a more reliable evaluation of the annotation accuracy.

Figure 3, shows the accuracy of the three developed annotation tools. The x-axis is the percentage of incorrectly assigned frames ($(1-OvR)\cdot 100\%$) and the y-axis is the percentage of phones that has a percentage of incorrectly assigned frames lower than the value given in the x-axis. The solid line represents the accuracy of the HMM-based aligner, the dashed line is the accuracy of the speech-synthesis-based aligner when it is used to refine the results of the HMM-based aligner. The dotted line represents the accuracy of the speech-synthesis-based

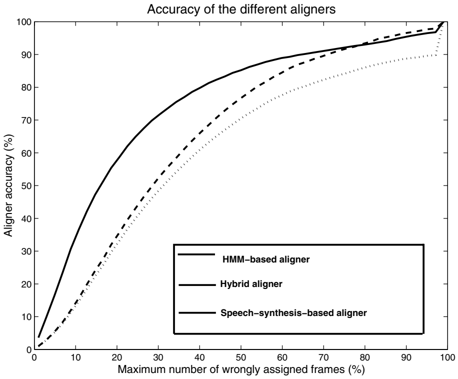

Fig. 3. Annotation accuracy for the three tested annotation techniques

aligner when no other alignments were available. In fact, these results are not a fair comparison among the multiple annotation tools, because the HMM-based aligner is an aligner adapted to the speaker, while the speech-synthesis-based aligners are not. Nevertheless, the phone models used in the HMM-based aligner were trained on data aligned by the the speech-synthesis-based aligner. These results also suggest that the use of HMM-based along with speech-synthesis-based annotation tools can be worthy as the former is more robust and the later is more accurate.

3 Confidence Scores

In this section we propose some phone-based confidence scores for detecting misalignments in the utterance. The goal is to locate regions of the speech signal where the alignment method may have failed and that could benefit from human intervention.

3.1 The Chosen Features

The alignment process provides a set of features that can be used as indicators of annotation mismatch. This set of features is described below.

- DTW mean distance: mean distance between the features of the recorded signal frames and the synthesized speech signal over the alignment path for a given phone;
- DTW variance: variance of the mean distance between the features of the recorded signal frames and the synthesized speech signal over the alignment path for a given phone;
- DTW minimal distance: minimal distance between the features of the recorded signal frames and the synthesized speech signal over the alignment path for a given phone;

- DTW maximal distance: maximal distance between the features of the recorded signal frames and the synthesized speech signal over the alignment path for a given phone;
- HMM mean distance: mean distance between the features of the recorded signal frames and the phone model;
- HMM variance: variance of the distance between the features of the recorded signal frames and the phone model;
- HMM minimal distance: minimal distance between the features of the recorded signal frames and phone model;
- HMM maximal distance: maximal distance between the features of the recorded signal frames and phone model

Each segment of the database is associated with a vector of features that will be used to predict a confidence score for the alignment of that phone. To provide some context we decided to include not only the feature vector of the current phone but also the feature vectors of the previous and following segments.

We were now in the condition of performing the evaluation the reliability of the different techniques that we propose to detect annotation problems.

Three different approaches will be evaluated: Classification and Regression Trees (CART), Artificial Neural Networks (ANN) and Hidden Markov Models (HMM).

3.2 Definition of Bad Alignment

A boundary between good and bad alignment is hard to define. Some researchers assume that boundary errors larger than 10 miliseconds must be considered misalignments, while others are more tolerant. As we explained before, the effect of the error in the location of the boundaries may be different from segment to segment, depending on its duration. Thus, we will use the duration-independent feature proposed before to computed the accuracy of annotation tools: we will assume that a misalignment occurs when $OvR \leq 0.75$.

3.3 Classification and Regression Trees

To train a regression tree we have used the *Wagon* program, that is part of Edinburgh Speech Tools[7]. This program can be used to build both classification and regression trees, but in this problem it was used as a regression tool to predict the values of the OvR based on the former features. We used a training set with 28000 segments and a test set with 10000 segments.

Since the leafs of the tree are the average value of OvR and its variance, assuming a gaussian distribution in the leafs, we can compute the probability of the having OvR with a value lower than the threshold defined in the previous subsection. Let μ and σ be the average value of OvR and its standard deviation, respectively, in a given leaf of the tree. Then, the probability of misalignment is given by:

$$P(OvR \leq 0.75|\mu,\sigma) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma^2}}\int_0^{0.75} e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}dx \tag{3}$$

We than had to apply a threshold to the resulting probability. By varying these threshold we obtained a Precision/Recall curve represented as a dotted line in Fig. 4.

3.4 Artificial Neural Networks

Using a neural network simulator developed at our lab, and the same feature vectors used in the previous experiment, we trained a binary classifier, which computes the probability of misalignment for each segment. As we did in the trainning of the regression tree, we had to apply a threshold to the outputs of the neural network. The variation of this threshold created the lower dashed line of Fig. 4.

3.5 Hidden Markov Models

Two one-state models were created for each phonetic segment. A model for aligned segments, and a model for the misaligned ones. Since the amount of training data were not large enough to build context dependent models, we had to choose a context-independent approach. However, we took into account the influence of the different contexts in the models in some extent by using four gaussian mixtures in each state. Each model was based on the feature vectors described in 3.1. After model training, we performed a forced alignment between the feature vector sequences and the model pairs trained for each phonetic segment. This experiment allowed us to find values for precision and recall for each phonetic segment. We depict the experiment results based on phone groups (Vowels, Liquids, Nasals, Plosives, Fricatives, Semi-Vowels and the Silence), which is enough to show that the precision and recall values can vary largely with the phone types in analysis.

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

Class	Precision(%)	Recall(%)
Vowels	73.2	69.8
Liquids	48.6	64.0
Nasals	82.0	67.7
Plosives	78.7	72.4
Fricatives	88.0	69.0
Semi-Vowels	44.9	67.5
Silence	97.3	87.8

Based on the previously trained models, we computed a score($HmmSore$) for each segment to precision-recall curves, like we did for CART and ANN. This score was calculated using equation 4.

$$HmmScore = \frac{P(x = Al|Model_{Al})}{P(x = Al|Model_{Al}) + P(x = Misal|Model_{Misal})} \quad (4)$$

where $P(x = Al|Model_{Al})$ is the probability that segment x is aligned given the model of aligned phones for that segment and $P(x = Misal|Model_{Misal}))$ is the probability that segment x is misaligned given its model of misaligned phones. The score values are between 0 and 1. We computed the upper curve of Fig. 4 by imposing different thresholds to the score, like we had already done for the two other approaches. It is important to point out that in this case we are detecting the **aligned** segments rather than **misaligned** ones.

3.6 Results

The results depicted in Fig. 4 suggest the HMM approach outperforms all others by far. The other two approaches are very similar, for some applications one should choose CARTs, for others one should choose ANNs.

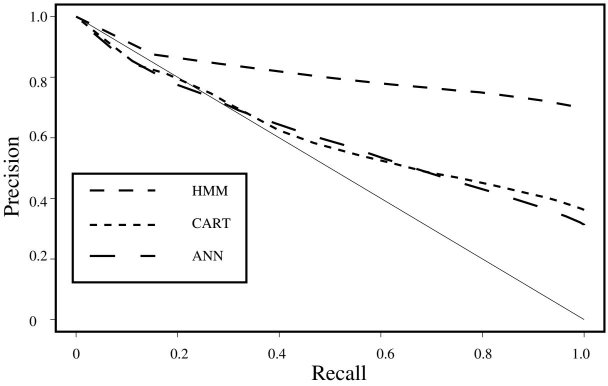

Fig. 4. Plot of precision and recall of the proposed confidence measures

4 Conclusions

In the first part of the paper, we have explored the advantages of using an HMM-based aligner together with an aligner based on speech-synthesis, and we showed the increase of the accuracy of the combined system, and a new measure of alignment accuracy was proposed. In the second part of the paper we proposed and evaluated three new approaches to compute confidence measures for phonetic annotation. In this part we realized that the approach using HMMs is largely the best one.

5 Acknowledgements

The authors would like to thank M. Céu Viana and H. Moniz for providing the manually aligned reference corpus. This work is part of Sérgio Paulo's PhD

Thesis sponsored by a Portuguese Foundation for Science and Technology (FCT) scholarship. INESC-ID Lisboa had support from the POSI Program.

References

1. D. Caseiro, H. Meinedo, A. Serralheiro, I. Trancoso and J. Neto, *Spoken Book alignment using WFST* HLT 2002 Human Language Technology Conference.
2. F. Malfrère and T. Dutoit, *High-Quality Speech Synthesis for Phonetic Speech Segmentation.* In Proceedings of Eurospeech'97, Rhodes, Greece, 1997.
3. Sakoe H. and Chiba,*Dynamic programing algorithm optimization for spoken word recognition.* IEEE Trans. on ASSP, 26(1):43-49, 1978.
4. J. Kominek and A. Black, *Evaluating and correcting phoneme segmentation for unit selection synthesis.* In Proceedings of Eurospeech'2003, Geneve, Switzerland, 2003.
5. A. Black, P. Taylor and R. Caley, *The Festival Speech Synthesis System.* System documentation Edition 1.4, for Festival Version 1.4.0, 17th June 1999.
6. S. Paulo and L. C. Oliveira, *DTW-based phonetic alignment using multiple acoustic features.* In Proceedings of Eurospeech'2003, Geneve, Switzerland, 2003.
7. P. Taylor R. Caley, A. Black, S. King, *Edinburgh Speech Tools Library* System Documentation Edition 1.2, 15th June 1999.