EXPLICIT SEGMENTATION OF SPEECH USING GAUSSIAN MODELS

Antonio Bonafonte, Albino Nogueiras and Antonio Rodriguez-Garrido
{antonio,albino}@gps.tsc.upc.es

Universitat Politècnica de Catalunya
c/Gran Capità s/n
08034 Barcelona (SPAIN)

ABSTRACT

In this paper we investigate an automatic method to segment labeled speech. The method needs an initial estimation of the segmentation which is provided by an alignment based on HMM. Afterwards, the boundaries are refined moving the frontier frames to the segment which is *more similar* to the speech frame. Gaussian *pdf* are used as a similarity measure. The performance of the method is evaluated using the TIMIT database. If boundary deviation (from the reference position) larger than 20 ms. are counted as errors, then the replacement of the boundaries reduces the error in a 30%. Additional experiments show how the proposed method turns the performance quite independent of the speaker dependent or speaker independent data used to estimate the HMM.

1. INTRODUCTION

Many speech processing problems require the labeling of a large amount of speech. For instance, the analysis of speech sounds, perhaps in different contexts, requires a tedious work of collecting samples of the sounds. Speech segmentation is also important in speech recognizers. Although actual techniques do not require an explicit segmentation, it is accepted that faster training and better results are achieved if the modeling techniques are initialized with some segmentation of the speech. This is specially important for the methods which require time consuming procedures for training, for instance, methods based on conexionist models. Furthermore, the boundary labels can be used to evaluate the recognition systems. The performance of the systems is usually measured aligning the recognized transcription with the reference transcription. However, frequently, specially when the error rate is high as for acoustic-phonetic decoding, the alignment count as corrects some errors of the recognizer. More coherent results are obtained if temporal information is used when comparing the reference and the test. Finally, segmentation of speech is also important in speech synthesis. There, a large number of units has to be segmented to buil a dictionary of subword units. This last case is the main motivation of this work.

In all the above mentioned problems, an orthographic annotation of the speech is available. Some works have appeared which find, not only the boundaries of the units, but also the uttered transcription which can be chosen from a set of proposed transcriptions. In Spanish, this problem is not so important as in other languages. The standard transcription can be derived automatically from the orthographic text. Although in some cases the phonetic transcription of the utterance differs from the standard transcription, mainly due to dialectal variations, this problem will not be considered on this paper. In the case of collecting elementary units for synthesis, the speaker is chosen to be *standard*.

Therefore, the paper treats with finding phonetic boundaries of a text given its phonetic content. This problem is known on the literature as *linguistically constrained segmentation* or *explicit segmentation*.

Most of the current speech segmentation techniques are based on modeling each one of the phonetic units, usually with Hidden Markov Models (HMM). The parameters of each HMM are estimated from a large number of samples of the phonetic unit, usually from many different speakers and from many contexts. This training material makes that HMM are able to represent accurately most of the situations being very suitable for recognition. However, in segmentation, as the phonetic content is known, the generalization of the models can be a drawback. It would be better to have more specific models, for instance, speaker dependent models, and word dependent models. This specific models could be adapted to the local properties of the speech. Unfortunately, usually there is not enough training data to estimate such models.

Some works have appeared which combine the explicit boundaries (obtained with HMM) and implicit boundaries (obtained directly from the speech signal) [2,4]. Implicit methods usually can be tuned to control the number of boundaries detected. If the number of boundaries is small, only some explicit boundaries can be improved by the implicit ones. On the other hand, it the number is too large, only very small movements are produced. The implicit method of [2] is based on the spectral changes but it reports no significant improvement over the use of only HMM. In [5], the implicit methods proposed in [2,4] do not produce any improvement with respect the use of HMM.

In this paper we propose a refinement of the segments based on the homogeneity of the speech segments. The hypothesis is that the speech frames of an acoustic phone have to be more similar to that phone than to the context phones. This hypothesis is more feasible for some sounds (as vowels) than for others (as plosive sounds), therefore, we pretended to apply the method only to some phonetic classes. However, as it will be shown on the results, it is not easy to determine which phonetic classes should be avoided. This idea was applied successfully to image segmentation [6], but there, as the regions where not known *a*

This research was supported by CICYT under contract TIC95-10220-C05-04

priori, a component of the model was defined to prevent oversegmentation.

The paper is organized as follows: next section presents the general segmentation scheme. Afterwards, the criterion to be used to measure the homogeneity of the segments is proposed. Section 4 presents some results with the TIMIT database. As the database has been hand labeled, reliable evaluation can be performed. Finally, section 5 analyzes the results for a Spanish database. In this results a standard phonetic transcription is used. The method is evaluated with different initializations, obtained from the use of speaker dependent and speaker independent HMM.

2. SEGMENTATION ALGORITHM

As it was stated on the introduction, first, a initial estimation of the boundaries is estimated. Afterwards, a procedure correct the boundaries position based on a homogeneity criterion.

The initial estimation is obtained using the Viterbi algorithm: given a set of HMM, one for each phonetic unit, the utterances to be segmented are mapped against a network composed by a HMM sequence. The HMM sequence is created from the phonetic transcription of the sentence which we assume it is known. The Viterbi algorithm finds the best path on the network. The backtracking of the path gives the initial segmentation.

Estimate a model e_i for each segment of the utterance *(i = 1…N)*
Compute the initial reference value $p_r = p(O_1 \ldots O_T / e_1 \ldots e_N)$
RepeatUntil the convergence
  ForEach boundary b_j *(j = 1…N-1)*
    Hypotheses 1: Move boundary b_j to the left
      Actualize models $\hat{e}_j^1$ and $\hat{e}_{j+1}^1$
      Compute $p_1 = p(O_1 \ldots O_T / e_1 \ldots \hat{e}_j^1 \hat{e}_{j+1}^1 \ldots e_N)$
    Hypotheses 2: Move boundary b_j to the right
      Actualize models $\hat{e}_j^2$ and $\hat{e}_{j+1}^2$
      Compute $p_2 = p(O_1 \ldots O_T / e_1 \ldots \hat{e}_j^2 \hat{e}_{j+1}^2 \ldots e_N)$
    If $max\{p_r, p_1, p_2\}$ ***is*** p_1,
      Move boundary b_j to the left
      $p_r = p_1$
      $e_j = \hat{e}_j^1$; $e_{j+1} = \hat{e}_{j+1}^1$
    If $max\{p_r, p_1, p_2\}$ ***is*** p_2,
      Move boundary b_j to the right
      $p_r = p_2$
      $e_j = \hat{e}_j^2$; $e_{j+1} = \hat{e}_{j+1}^2$
    EndIf
  End ForEach
End RepeatUntil

Figure 1: Algorithm used to refine the position of boundaries.

If several transcriptions were allowed, the network would be slightly more complicated and the backtracking would produce, not only the segmentation, but also the most probable transcription given the utterance.

The corrective procedure is presented in figure 1. It is applied independently to each utterance to be segmented. For each segment of the utterance, a model is estimated. If one particular phonetic unit appears *n* times on the utterance then *n* models are estimated for that unit. The join probability of the utterance given the models is computed and taken as reference. Then, for each boundary, the hypotheses of moving the boundary one frame to the left or one frame to the right is analyzed: for each movement, the join probability is computed and, if the value is higher than the reference, the boundary is moved and the models and the probability reference is actualized. The algorithm is iterated until no movement produces an increase on the join probability.

3. THE HOMOGENEITY CRITERION

In last section an algorithm has been presented which estimate the boundaries between phones using local models of the segments. The first idea was to use a single Gaussian *pdf* to model the feature vector (12 mel-cepstrum coefficients). However, four variants have been tried to avoid estimation problems on the covariance matrix in those segments with few data:

1. **GAU**: the covariance matrix of the models is assumed to be diagonal. For each segment the mean and the covariance are estimated from the data of that segment.
2. **MAH**: the covariance is estimated using the data of the whole utterance.
3. **LIG**: the covariance matrix of the model is assumed to be the identity matrix weighted by a factor. This approximation implies that the models have spherical symmetry.
4. **EUC**: the covariance matrix of the models is assumed to be the identity. This approximation implies that the models have spherical symmetry with the same ratio for all the models. In fact, this model is computed using the Euclidean distance.

Note that using Gaussian *pdf* to characterize the segments, the models can be adapted very easily each time that a hypothesis is considered. This makes negligible the computational cost of the algorithm.

As it was mentioned in the introduction, the algorithm assumes that phones are stationaries, which is far from being true for some sounds. In order to cope with this limitation, the algorithm has been modified so that the homogeneity criterion depends on the type of sound. Particularly, the acoustic labels of the TIMIT have been grouped on 11 phonetic classes: vowels, closures, nasals, diphthongs, etc. The algorithm proposed above was applied to the utterances which were used for training the HMM and, for each pair of phonetic classes $\{C_i, C_j\}$, it was determined which one of the above criteria produced the best position of the boundary

between the two sounds. One criterion was assigned during training to each $\{C_i, C_j\}$, being one of the possible criteria not to refine the boundary. In this way, refinement is applied only on boundaries between sounds where the stationary hypothesis is appropriated. We have named this criterion as **MIX**. It should be noted, that this criterion requires a large labeled database and therefore, it can not be applied to those databases where such information is not available.

4. EVALUATION USING THE TIMIT DATABASE

The first evaluation of the algorithm has been performed using the TIMIT database [6] which provides enough data to obtain results with statistical significance.

HMM are trained with 1,600 utterances, selected from the 3,696 utterances of the training corpus. At least one sentence of each speaker of the training corpus has been selected. The results which are presented have been obtained using the 192 sentences of the core test set. Similar results are obtained if the whole test set is used.

An acoustic vector is formed each 10 milliseconds (limiting the resolution of the method) by analyzing speech frames of 20 milliseconds using a Hamming window. For each frame, the output of 20 mel-scaled filters is transformed on 12 mel cepstrum coefficients. Furthermore, the first and the second derivative, and also the first power derivative is used by the HM Models. However, only the 12 cepstra coefficients are used to refine the boundaries. The HMM have four states and the Bakis topology: therefore, the minimum length of the segments to be detected is 30 milliseconds. The output probabilities densities are modeled by semicontinuous models using 512 gaussians for the cepstrum, 256 for the cepstrum derivatives and 128 for the power derivative. The output probabilities are approximated by the *M* most relevant contributions, being *M* equal to 6 for cepstrum coefficients and 2 for the power coefficient.

Table 1 shows, for each method, the percentage of boundaries which present an error smaller than 20 milliseconds (first row) or 12 millisecond (second row). In the same table, the letters *INI* are used to indicate the values achieved if the method is not applied, it is, if the sentences are segmented using only HMM.

It can be seen how all the methods reduce the error with respect to the use of only HMM. The best results are obtained using the Euclidean distance to the mean of the segment. In this case, the error is reduced around a 30%.

t	*INI*	*GAU*	*LIG*	*MAH*	*EUC*	*MIX*
20 ms	73.6	75.6	77.6	79.5	79.9	81.3
12 ms	52.7	55.6	59.4	62.2	63.4	64.4

Table 1: Percentage of boundaries which have been detected with an error of less than a) 20 ms., b) 10 ms., as a function of the model used as homogeneity criterion.

	VOW	**CLO**	**SEM**	**NAS**	**UST**	**VST**	**UFR**	**VFR**	**SIL**	**DIP**	**AFF**
VOW	EUC	EUC	GLB	EUC	GLB	INI	EUC	EUC	LIG	EUC	INI
CLO	EUC	INI	LIG	LIG	GLB	EUC	GAU	GLB	INI	INI	EUC
SEM	GLB	EUC	EUC	EUC	GLB	EUC	EUC	GLB	EUC	EUC	INI
NAS	GLB	INI	EUC	INI	EUC	INI	EUC	EUC	GLB	EUC	EUC
UST	INI	INI	LIG	GLB	GLB	EUC	EUC	EUC	INI	INI	INI
VST	INI	INI	EUC	EUC	INI	NA	GAU	INI	LIG	EUC	NA
UFR	EUC	INI	EUC	LIG	INI	NA	INI	INI	EUC	EUC	INI
VFR	GAU	INI	GLB	GAU	GAU	INI	GAU	GLB	INI	GAU	NA
SIL	EUC	GLB	EUC	EUC	EUC	EUC	EUC	EUC	INI	EUC	EUC
DIP	INI	INI	EUC	INI	EUC	INI	EUC	EUC	INI	INI	NA
AFF	EUC	INI	LIG	EUC	INI	NA	INI	EUC	GAU	EUC	NA

Table 2: Method which gives better performance (over the training utterances) with respect to the neighbor acoustic class which define the boundary. *(vow: vowels; clo: closures; sem: semivowels; na: nasals; {u|v}st: {unvoiced|voiced} stops; {u|v}fr: {unvoiced|voiced} fricative; sil: silence; dip: diphthongs; aff: affricates)*. Columns: previous sounds; row: posterior sound.

The **MIX** criterion, which applies different criteria to each pair of sounds gives slightly better results. Table 2 shows, for each pair of acoustic classes, the method which has been assigned. For instance, it can be observed how the boundaries between any sound and closures, only should be refined it the first sound is a vowel or semivowel (apart from silence). The number of sound combinations which should not be refined is 35 from a total number of 114.

Figure 1 shows the histogram of boundaries deviations from true boundary positions. In order to see the performance of the different criteria, the histogram is plotted as a continuous line, as a estimation of the distribution probability of the deviation. It can be observed how the method, for all the criteria, removes deviation errors of around 20 milliseconds. It is noticeable the increase of the number of boundaries which are detected exactly as in the original labels.

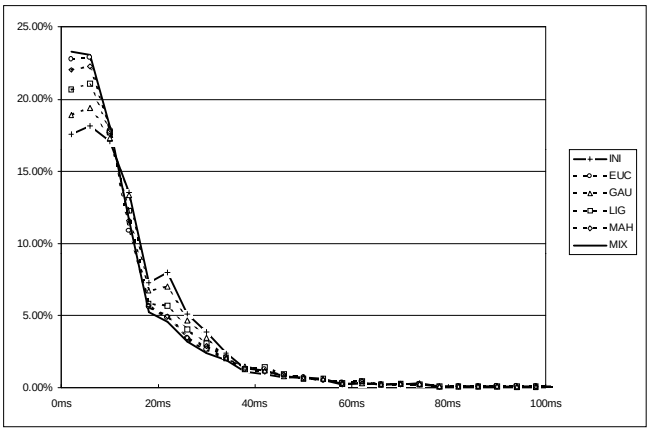

Figure 1: histogram of boundary placements errors for different homogeneity criterion.

4. EVALUATION USING A SPANISH DATABASE

The previous results have given the evidence that the proposed method can be used to improve the segmentation results produced by HMM segmentation. In this section, the method is tested on a Spanish database which has been recorded by the *Universidad Politécnica de Valencia*. The database consist of 170 sentences which are repeated by 10 speakers. Hand labels are available for 77 utterances.

The aim of the experiments is to evaluate the method when it is used to detect phone boundaries. Furthermore, the phonetic transcription is derived automatically from the orthographic text. Finally, the experiments analyze the convenience of using speaker dependent models with generic speaker independent models. As our purpose is to segment logatomes to extract subword units for a text to speech system, if speaker dependent models were used, a speaker dependent database should be recorded. It would be better if generic models could be used.

Three experiments has been done where the first initialization is performed with HMM which are *a)* speaker dependent (SD), *b)* multi-speaker (MS) and *c)* speaker independent (SI). The number of training utterances is around 170, 10×170 and 9×170 for each case. In the SI case, a *leaving one out* strategy was followed so that the 77 hand labeled utterances can be used for testing the system. The low number of labeled utterances diminishes the significance of the results.

From the HMM initialization, the segmentation algorithm of section II has been applied using Euclidean distance. The results given on table 3 are better than those obtained for the TIMIT database.

Results show how, if only HMM are used, it is better to estimate speaker dependent models, even in the case that the number of training sentences is much smaller. The multispeaker seems a good trade off because more training data is available and the models are in some way adapted to the speaker. On the other hand, when boundaries are refined, in the three cases, the final location becomes closer to the reference ones. Furthermore, the importance of the initial segmentation is reduced: almost the same results are obtained with the different initializations. Therefore, the use of generic speaker independent HMM for the first step of the algorithm does not degrades the performance.

	MS	***SI***	***SD***
INI	84	80	82
EUC	86	85	84

Table 3: Percentage of boundaries which have been detected with an error of less than 20 ms. as a function of the model used as homogeneity criterion, and the data used to train the HMM.

5. CONCLUSIONS

In this paper a method has been proposed to segment the speech. The method considers the segments of segmented speech as homogeneous regions. Therefore, a homogeneity criterion is applied to refine the frontiers of the regions. The method needs an initial estimation of the boundaries which is provided using a HMM based segmentation scheme. Different proposals are tried to define the homogeneity criterion; in fact the criterion is the assumption that the cesptra coefficients of the segments can be properly characterized by a Gaussian *pdf* and the different variations is the assumptions assumed to estimate the parameters of the *pdf* to cope with sparse data. The results show how the method reduces 30% the number of boundaries whose placement error is larger than 20 ms. Some additional experiments show how the method is in some way independent of the exact of the initialization (as far as it is sufficiently enough). Thus, the results are similar if speaker dependent or speaker independent HMM are used to the produce the first initialization.

6. REFERENCES

1. A. Marzal and E. Vidal, "A review and new approaches for automatic segmentation of speech signals", *Proc. of EUSIPCO'90*, pp. 43-53, Barcelona, 1990.

2. F. Brugnara, D. Falavigna and M. Omologo, "Automatic segmentation of speech based on Hidden Markov Models", *Speech Communication*, No 12, pp. 357-370, 1993.

3. A. Ljolie and M.D. Riley, "Automatic segmentation and labeling of speech", *Proc. of ICASSP'91*, pp. 473-476, Toronto, 1991.

4. J.P. van Hermert, "Automatic Segmentation of Speech", *IEEE Trans. on Signal Processing*, No 4, pp. 1008-1012, April, 1991.

5. A. Rodriguez Garrido, *Segmentación de voz por métodos explicitos e implicitos*, Master Dissertation, ETS. d'Enginyers de Telecomunicació de Barcelona, 1995.

6. L. Lamel, R. Kassel and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", *Proc. of the DARPA Speech Recognition Workshop*, pp. 26-32. March, 1987.

7. F. Marqués, A. Gasull, T. Reed and M. Kunt, "Coding-oriented segmentation based on Gibbs-Markov random fields and human visual system knowledge", *Proc. of ICASSP'91*, pp. 2749-2752, Toronto, 1991.