AN ACOUSTIC SUBWORD UNIT APPROACH TO NON-LINGUISTIC SPEECH FEATURE IDENTIFICATION

Mohamed Afify[1] *Yifan Gong*[1,2] *Jean-Paul Haton*[1]

[1]CRIN/CNRS-INRIA-Lorraine,B.P. 239 54506 Vandeouvre,Nancy,France
[2] Media Technologies Laboratory, Texas Instruments, P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.

ABSTRACT

Automatic identification of non-linguistic speech features (e.g. the speaker or the language of an utterance) are currently of practical interest. In this paper, we first impose a set of requirements that we think a statistical model used in non-linguistic feature identification should satisfy. Namely, these requirements are capturing both short and long term correlations in addition to maintaining a certain acoustic resolution. A model satisfying these requirements, and in the same time having the attractive feature of requiring no transcribed speech material during training is proposed. Experimental evaluation of the approach in speaker recognition on the TIMIT database is presented, where recognition rates up to 99.2 % are achieved.

1. INTRODUCTION

Automatic identification of non-linguistic speech features [1] (e.g. speaker identification or language identification) are currently of interest in many practical systems. A key point in successful acoustic modelling for automatic non-linguistic feature recognition is that of characterizing the features's acoustic space with sufficient resolution and in the same time capturing the feature's specifities. For a given acoustic resolution (e.g. number of Gaussians) we distinguish different modelling approaches based on their ability of capturing short term (frame level) and long term (unit level)correlations, as both are assumed to be potentially discriminant in the identification process. Models proposed in the literature differ in their ability of statisfying these requirements.

For example statistical approaches in the field of automatic talker recognition can be classified according to this point of view. Frame based techniques that use a mixture of Gaussians [1], or a VQ codebook [2] are by principle unable to capture both types of correlations, while those utilizing a conventional frame level mixture distributed among different states (e.g. an ergodic HMM) [3] model only long term correlations and ignore the short term ones. On the other hand segment based techniques as matrix quantization methods [4, 5], orthogonal polynomial representation [6], and autoregressive models [7] aim at modelling only short term correlations. Also in the field of automatic language identification (see [12, 13] for a review) Gaussian mixtures don't satisfy both requirements. While the widely used approach of HMM tokenization followed by phonotactic modelling captures only long term correlations, and leads to the need of long speech intervals (on the order of 45 seconds) to perform accurate identification. In addition, the accuracy of using phone models of a certain language to tokenize another language is clearly questionable. [2] Moreover, both fields i.e automatic speaker and language identification are usually treated as separate problems, inspite of their strong resemblance from an acoustic modelling point of view.

Recently a realization satisfying the above point of view, in capturing both types of correlation, and pursuing a unified approach to non-linguistic feature identification, was reported in [8]. In this implementation, the acoustic space is divided into homogenous regions corresponding to phonemes, and each region is associated to the state of a large ergodic Markov model, and is characterized by a left to right continuous density hidden Markov model (CDHMM). Feature identification reduces to parsing a test utterance in a Viterbi sense, using all existing models and choosing the highest scoring model. However, model training in this approach requires a set of phonetically labelled speech material, in addition to speaker independent phonetic models for appropriate initialization. [3]

These requirements maybe a limitation in practical situations, e.g. for automatic language identification systems, where the flexibility of adding new languages, with possibly unknown phonetic sets, may be desirable.

In this paper we first identify the requirements that we think a statistical model used in non-linguistic feature identification must satisfy as:

- Maintain a certain acoustic resolution.
- Capture short term correlations (frame level).
- Capture long term correlations (unit level).
- Can be constructed without the need for transcribed speech material.

Then we present a realization satisfying these requirements, together with an experimental evaluation of the proposed method for speaker recognition on the TIMIT database.

[1]This term was proposed by Lamel and Gauvin in [8]

[2]An interesting approach to automatic language identification that also doesn't require transcribed speech was recently proposed in [14].

[3]As noted earlier the accuracy of using phone models of a language to tokenize another language is clearly questionable.

The paper is organized as follows. Section 2 gives a general overview of the proposed method. An automatic segmentation method related to system construction is reviewed in Section 3. Section 4 discusses stochastic trajectory models which are related to acoustic space characterization in our method. A probabilistic labelling strategy is introduced in Section 5. Sections 6 and 7 present the use of the model for identification, and its experimental evaluation. Finally we conclude in Section 8.

2. GENERAL OVERVIEW

To satisfy the above requirements we propose to model a feature space using a mixture of stochastic trajectories (of size K)[9, 10], this mixture maintains the required acoustic resolution and at the same time captures short term correlations. We then develop a probabilistic labelling strategy which, given the mixture model, assigns the training material to a set of acoustically homogenous regions corresponding to the mixture components. At this end, we have two alternative ways to charcterize the acoustic space for feature identification.

The first alternative is to use the mixture model itself to characterize the acoustic space. In this case the model is conceptually comparable to other segment based approaches (e.g. [4]-[7] in the field of speaker identification). This approach will be called the segmental approach (SEG). The second alternative consists in applying the probabilistic labelling strategy to assign the training data to K distinct acoustic regions. Then to construct an ergodic HMM, where each state corresponds to an acoustic region, and is characterized by a continuous density left to right HMM constructed from the corresponding training data. This model will potentially capture both types of correlations, i.e., short term correlations through the state HMMs and long term correlations through the ergodic Markov model. In addition it is constructed automatically from unlabelled training speech. This model will be referred to as the segmental HMM (SHMM).

In the light of the above discussion the basic ingredients of the proposed approach are: an automatic segementation procedure to initialize the stochastic trajectory model, the stochastic trajectory model, the probabilistic labelling strategy, and finally the use of the model for feature identification. Each of these components will be discussed in the following sections.

3. AUTOMATIC SEGMENTATION

In this section we will present an automatic segmentation procedure which is used to bootstrap the training process of the stochastic trajectory mixture model (discussed in the next section). The algorithm is due to Svendsen and Soong [11], and is briefly reviewed here.
For an utterance $\{x_1, ..., x_T\}$ of T frames the total distortion associated to a segmentation which consists of N_s segments can be written as:

$$D = \sum_{n=1}^{N_s} \sum_{t=t_{n-1}+1}^{t_n} d(x_t, C_n) \quad (1)$$

where x_t is the t^{th} frame of the n^{th} segment starting at frame $t_{n-1}+1$ and ending at t_n, where by definition $t_0 = 0$, and $t_{N_s} = T$, and C_n is the centroid of the n^{th} segment, which for the Euclidean distance used in this paper can be written as:

$$C_n = \frac{1}{t_n - t_{n-1}} \sum_{t=t_{n-1}+1}^{t_n} x_t \quad (2)$$

The objective of the segmentation algorithm is to assign segment boundaries to minimize the total distortion in (1). This can be efficiently achieved using dynamic programming(DP). The cumlative distortion $E(t)$ at frame t can be written as:

$$E(t) = \min_{d_{min} \le d \le d_{max}} \{E(t-d) + \sum_{i=t-d+1}^{t} d(x_i, C^t_{t-d+1})\} \quad (3)$$

where d_{min} and d_{max} are minimum and maximum duration constraints, and C^t_{t-d+1} is the centroid of the segment $\{x_{t-d+1}, ..., x_t\}$, and is calculated as in (2). When the end of the utterance is reached the optimal segmentation can be retreived using backtracking.

4. STOCHASTIC TRAJECTORY MIXTURES

A speech segment X_n of length d is modelled as being generated by a mixutre of K (mixutre size) trajectory generators, and we can write:

$$P(X_n|d) = \sum_{k=1}^{K} P_k P(X_n|k, d) \quad (4)$$

where P_k is the a priori probability of mixture component k, and each trajectory is assumed to consist of a set of Q(=5 for the current implementation) independent Gaussian states. Hence $P(X_n|k, d)$ can be written as:

$$P(X_n|k, d) = \prod_{i=0}^{Q-1} \mathcal{N}(x_{n,i}; \mu_{k,i}, \Sigma_{k,i}) \quad (5)$$

where $\mathcal{N}()$ is a normal distribution, with mean $\mu_{k,i}$, and covariance $\Sigma_{k,i}$(diagonal covariances are used in this work), and $x_{n,i}$'s are obtained by linear resampling of the segment X_n to the Q points of the trajectory.The model is charcterized by the parameter set $\lambda = \{P_k, \mu_{i,k}, \Sigma_{i,k}\}$ where $1 \le k \le K, 0 \le i \le Q-1$, and the apriori probabilities satisfy the stochastic constraint $\sum_{k=1}^{K} P_k = 1$.
In the following subsections we discuss the training process of the stochastic trajectory model.

4.1. EM estimation

Given a training set $X = \{X_n \;\; 1 \le n \le N\}$,the role of the training algorithm is to estimate the model parameter set λ to maximize the likelihood of the training data. This can be formulated as:

$$\lambda^* = \underset{\lambda}{\text{argmax}} \prod_{n=1}^{N} P(X_n|d_n, \lambda) \quad (6)$$

An efficient solution to this problem can be obtained using the expectation-maximization (EM) algorithm, and consists in applying the following EM steps:

1. E-step

$$P(k|X_n, d) = \frac{P(X_n|k, d)P_k}{\sum_{k=1}^{K} P(X_n|k, d)P_k} \quad (7)$$

2. M-step

$$P_k = \frac{1}{N}\sum_{n=1}^{N} P(k|X_n, d) \quad (8)$$

$$\mu_{i,k} = \frac{1}{N}\sum_{n=1}^{N} P(k|X_n, d)x_{n,i} \quad (9)$$

$$\Sigma_{i,k} = \frac{1}{N}\sum_{n=1}^{N} P(k|X_n, d)x_{n,i}x_{n,i}^T - \mu_{i,k}\mu_{i,k}^T \quad (10)$$

starting from an initial parameter estimate until convergence, where each iteration ensures the increase of the observed data likelihood.

4.2. Automatic resegmentation

Once the model parameters are obtained, they can be used to resegment the training material. For an utterance of length T the segmentation problem can be formulated as:

$$\mathcal{S}^* = \underset{\mathcal{S}}{\text{argmax}} \prod_{n=1}^{N_s} P(X_{t_{n-1}+1}^{t_n}|t_{n-1}, t_n, \lambda) \quad (11)$$

where $\mathcal{S} = \{t_1 = 0, t_2, ..., t_{N_s} = T-1\}$ is a segmentation of the utterance having N_s segments, and $X_{t_{n-1}+1}^{t_n}$ is used to denote $\{x_{t_{n-1}+1}, ...x_{t_n}\}$.
The above segmentation problem can be efficiently solved using dynamic programming (DP). The cumulative log probability at frame t ($L(t)$) is calculated as:

$$L(t) = \max_{d_{min} \le d \le d_{max}} \{L(t-d) + d * \log P(X_{t-d+1}^{t}|d)\} \quad (12)$$

where $0 \le t \le T-1$, T is the utterance length,and d_{min}, d_{max} are minimum and maximum duration constraints on the segment length, and the multiplication by d is a heuristic used to account for the resampling of the segment into a fixed length sequence. When the end of the utterance is reached the best segmentation can be retreived by using backtracking, if the best segement duration is kept for each time instant during the DP search.

4.3. Summary of the training algorithm

The complete training algorithm of the trajectory model consists of iterative application of the EM steps (7) and (8-10), and the resegmentation step (12) starting from appropriate initial model parameters. A summary of the whole algorithm is outlined below:

1. Perform initial automatic segmentation of the training speech as discussed in Section 3.
2. Using the results of the initial automatic segmentation, initialize the mixture components using LBG algorithm with distance measure (5), and ML parameter estimation.
3. Resegment the training speech using the mixture model by applying the DP algorithm in (12).
4. Apply the EM training step, by repeating the following two steps until convergence
   - For each training segment perform the E-step (7).
   - Estimate the model parameters using the M-step (8-10).
5. If convergence is not met go to step 3.

5. PROBABILISTIC LABELLING AND MARKOV MODEL CONSTRUCTION

The trajectory mixture model can be used to assign the training speech to acoustically homogenous regions, which we refer to as probabilistic labelling. This can be formulated as:

$$(\mathcal{S}, \mathcal{K})^* = \underset{(\mathcal{S},\mathcal{K})}{\text{argmax}} \prod_{n=1}^{N_s} P(X_{t_{n-1}+1}^{t_n}|t_{n-1}, t_n, k_n, \lambda) \quad (13)$$

where $\mathcal{S}$ is a segmentation as defined above, and $\mathcal{K}$ is a set of mixture labels.
Again as in the case of the segmentation problem this labelling strategy can be efficiently solved using dynamic programming (DP). Let $L(t)$ be the cumulative log probability score at frame t, and d_{min}(=2 for the current implementation) and d_{max}(=20 for the current implementation) be minimum and maximum segemnt duration constraints respectively. We can write:

$$L(t) = \max_{1 \le k \le K} \max_{d_{min} \le d \le d_{max}} \{L(t-d) + d * \log P(X_{t-d+1}^{t}|k, d)\} \quad (14)$$

By keeping the best duration and label index at each time frame, we can backtrack the best label sequence once we reach the end of the utterance.

After probabilistic labelling of the training data, all segments corresponding to a mixture label are assigned to the same acoustic region, and are used to construct a left to right HMM for that region. This HMM construction process can be performed using classical Baum-Welch training. In addition the bigram frequencies of the labels are used to estimate the transition probabilities of an ergodic HMM, where each state of this model is characterized by the corresponding region HMM. Further refinement of the model parameters can be done by iterating segmentation of the training data and model estimation until convergence. However, this was not done in this paper. The model construction process is summarized below.

1. Perform probabilistic labelling of the training data. By applying the DP recursion (14) and backtracking.
2. Assign segments having the same label index to the same acoustic region.
3. For each acoustic region construct a left to right CDHMM.
4. Assign each region to a state of an ergodic HMM, and estimate the trasition probabilities of this model as bigram frequencies of the labels.

6. FEATURE IDENTIFICATION

A model (whether segmental or segmental HMM) is constructed for each non-linguistic feature of interest. Non-linguistic feature identification includes parsing an unknown test utterance using all existing models, and choosing the feature yielding the highest probability score. For the trajectory mixture model the parsing step uses DP which is similar to (12), while for the Ergodic HMM it utilizes classical Viterbi decoding. Pruning can be optionally introduced for both models to save computation.

7. EXPERIMENTAL RESULTS

We evaluate the proposed approach for speaker recognition on TIMIT. 114 speakers corresponding to the first two dialect regions are used in the test. 8 sentences (sx+si) from each speaker are used for training, and 2 sentences (sa) are used for test (2*114 = 228 test utterances). 12 MFCC are used for speech parametrization. We give results for both trajectory mixture (SEG) and segmental HMM with state HMMs having 1 mixture component/state (SHMM1), and 2 mixture components/state (SHMM2). The results are shown in the table below. The presented results show the efficacy of the approach. For the presented results the segmental HMM approach outperforms the segmental one. However, further experimentation is needed to justify the results.

Mixture size	SEG	SHMM2	SHMM1
4	87.3	98.3	97.8
8	96.5	99.2	98.3

Table 1. Percent accuracy speaker recognition results on the TIMIT database.

8. CONCLUSION

We have presented a unified statistical approach to non-linguistic speech feature identification. The approach is based on charcterizing the acoustic space with a mixture of stochastic trajectories. Based on this representation two distinct approaches to non-linguistic feature identification were discussed. The first is based on the stochastic trajectory mixture, while the other uses a probabilistic labelling strategy to charcterize the feature space with an ergodic Markov chain. The latter approach captures both short term and long term correlations, and in the same time doesn't require transcribed speech material for training. In the experimental evaluation the ergodic model appraoch outperformed the mixture trajectory one. However, further experimentation is still needed for better justification. Also we plan to apply the proposed methods to the problem of automatic language identification.

REFERENCES

[1] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech communication, vol. 17,pp.91-108, August 1995.

[2] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang, " A vector quantization approach to speaker recognition," AT&T Technical Journal, Vol. 66, pp.14-26, Mar/Apr. 1987.

[3] T. Matsui, and S.Furui, "Comparison of text independent speaker recognition methods using VQ-distortion and discrete continuous HMM's", IEEE Trans. Speech and Audio Processing,vol. 2, No.3, pp.456-459, July 1994.

[4] M.S. Chen, P.H. Lin, and H.C. Wang, "Speaker identification based on a matrix quantization method," IEEE Trans. Signal Processing, vol. 41, No. 1, pp. 398-403, Jan. 1993.

[5] B.H. Juang, and F.K. Soong, "Speaker recognition based on source coding approaches," Proc.ICASSP-90, pp.613-616.

[6] C.S. Liu, H.C. Wang, F.K. Soong, and C.S. Huang, "An orthogonal polynomial representation of speech signals and its probabilistic model for text independent speaker verification," Proc. ICASSP-95, pp.345-348, May 1995.

[7] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet, "Standard and target driven AR vector models for speech analysis and speaker recognition," Proc. ICASSP-92, pp. II.5-II.8, March 1992.

[8] L. Lamel, and J.L. Gauvin, "A phone based approach to non linguistic speech feature identification," Computer speech and language, vol. 9, No.1, pp.87-103, Jan. 1995.

[9] Y. Gong, and J.P. Haton, "Stochastic trajectory modelling for speech recognition,"Proc. ICASSP-94, pp.57-60, April 1994.

[10] Y. Gong, "Stochastic trajectory modeling and sentence searching for continuous speech recognition",IEEE Trans. on Speech and Audio Processing, vol.5, No.1, pp.33–44, Jan. 1997.

[11] T. Svendsen, and F. Soong, "On the automatic segmentation of speech signals," Proc. ICASSP-87, pp.77-80, April 1987.

[12] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech and Audio processing, Vol.4, No. 1, pp. 31-44, Jan. 1996.

[13] Y.K. Muthusamy, E. Barnard, and R. Cole, "Reviewing automatic language identification," IEEE Signal Processing Mag., Vol. 11, No. 4, pp. 33-41, Oct. 1994.

[14] M.A. Lund, K. Ma, and H. Gish, "Statistical language identification based on untranscribed training," Proc. ICASSP-96, Vol. 2, pp. 793-796.

AN ACOUSTIC SUBWORD UNIT APPROACH TO NON-LINGUISTIC SPEECH FEATURE IDENTIFICATION

Mohamed Afify[1] , *Yifan Gong*[1,2] *and Jean-Paul Haton*[1]

[1]CRIN/CNRS-INRIA-Lorraine,B.P. 239 54506 Vandeouvre,Nancy,France

[2] Media Technologies Laboratory, Texas Instruments, P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.

Automatic identification of non-linguistic speech features (e.g. the speaker or the language of an utterance) are currently of practical interest. In this paper, we first impose a set of requirements that we think a statistical model used in non-linguistic feature identification should satisfy. Namely, these requirements are capturing both short and long term correlations in addition to maintaining a certain acoustic resolution. A model satisfying these requirements, and in the same time having the attractive feature of requiring no transcribed speech material during training is proposed. Experimental evaluation of the approach in speaker recognition on the TIMIT database is presented, where recognition rates up to 99.2 % are achieved.