ISCA Archive
http://www.isca-speech.org/archive

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Speaker Verification Based on Broad Phonetic Categories

Sachin S. Kajarekar[1], *Hynek Hermansky*[1,2]

[1] Oregon Graduate Institute of Science and Technology, OR, USA
[2] International Computer Science Institute, Berkeley, CA, USA
sachin@asp.ogi.edu, hynek@asp.ogi.edu

Abstract

In this work we present a speaker verification system based on 4 broad phonetic categories: vowels+diphthongs, fricatives, glides+nasals, and silence+stops. Using these categories separately, it is observed that vowels, diphthongs, and fricatives are the most important categories for speaker verification. This observation confirms the results from the analysis of speaker and channel variability in speech. Using NIST speaker verification evaluation data, the performance of the phone based system is compared with the conventional speaker verification system based on Gaussian mixture model (GMM). The results show that the phone-based system outperforms the conventional system specifically when there is channel mismatch between training and testing data.

1. Introduction

Speaker verification is a process of verifying the identity of the speaker based on his/her speech. In text-independent speaker verification, different text can be used for training and testing. Phone-based speaker verification is performed in many stages. In the first stage, a sequence of phones is estimated from a given utterance using a speech recognition system. Then speaker verification is performed separately for each phone to obtain a verification score. The score for each phone is weighted according to the importance of the phone for speaker verification. The final verification result is obtained by combining the weighted results from all phones.

The main problem with the phone-based system has been that, the sequence of phones can not be accurately estimated using a simple speech recognition system. This problem has been addressed in several ways. Some researchers [1, 2, 3] have used a small set of broad phonetic categories assuming that phones with a category are equally important for speaker verification. Others [4] have used a complex state-of-the art speech recognition system. In this work, we have used the first approach.

Among the previous work, Gupta and Savic [1] have used four broad phonetic categories - voiced, fricative, nasal, and plosive - in their system. The speech recognition was performed in two steps. First, N state auto-regressive hidden Markov model (HMM) was trained using the input utterance and the trained HMM was used to segment the utterance into different states. The state means were then associated with one of the broad phone categories using rule-based approach. The speaker verification was performed using class-specific features and final score was obtained by a weighted combination of results obtained using vowels, fricatives and nasals. It was concluded that for the speaker recognition task, plosives are least effective whereas the voiced phones and fricatives are the most effective broad classes. Same microphone was used for training and testing in this work. The size of the database was also rather small.

Parris and Carey [2] used HMMs for speaker verification too. They showed that the log-likelihood ratio can be used to characterize speaker discriminating ability of phones. In their system, speech recognition was performed using speaker independent sub-word models. The log-likelihood ratio between the speaker independent model and the claimed speaker model was used for speaker verification. It was concluded that a subset of phones - front vowels, voiced fricatives and nasals - outperforms the complete set of phones on the text-independent speaker recognition task. However, the database used for these experiments was collected over the same telephone line and the size of the test database was rather small.

Koolwaaij and de Veth [3] reported speaker verification results on NIST 1998 evaluation data that is larger than the ones used in previous works. They used vowel, fricative, plosive, nasal, liquid and silence as broad categories. These categories were obtained using a speech recognition system trained on an independent database. Speaker verification results were obtained independently for each category. The difference in the performance for broad categories was attributed to their longer duration and frequent occurrence. The combination of the proposed system and the baseline system was shown to give a significant improvement. However, the comparison of the system with a state-of-the-art speaker verification system [5] was not done.

In our previous work [8] the phone specific speaker and channel variability was studied using analysis of variance (ANOVA). We showed that vowels, diphthongs, nasals and fricatives are the most important categories for speaker verification. In this work we verify these results using a speaker verification system based on 4 broad phonetic categories. Section 2 describes the results of the analysis of variance (ANOVA). Section 3 describes two SPVER systems : 1) baseline system using Gaussian mixture models (GMMs) [6], and 2) the phone-based system using HMMs. Section 4 describes the database and features used in these experiments. The results are described in Section 5 which show that the phone based system outperforms the GMM-based system with the same model complexity. We conclude the paper with the summary of results in section 6.

2. Estimation of phone specific speaker and channel variability

Speech is primarily used to convey linguistic information. However, it also contains other sources of information like speaker and communication channel (telephone handset and telephone line). Here linguistic information refers to variability in the language due to different phones. Speaker and channel information refer to variation in speech signal introduced by different speaker and channel characteristics respectively. In our previous work [8], we used ANOVA to estimate the speaker and channel variability for different phones. Figure 1 shows the results of the analysis for the broad categories. It shows that vowels, diphthongs, nasals and fricatives contain the highest amount of speaker variability. Stops show the lowest amount of variability for different speakers. In this work, we have verified these results using speaker verification experiments. The verification systems used in the experiments are described in the following section.

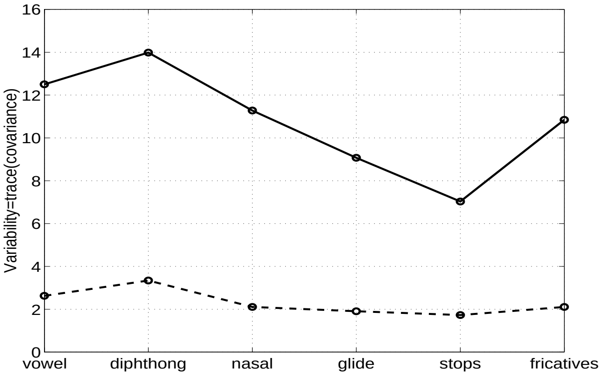

Figure 1: The phone-specific speaker and channel variability. Solid line represents the speaker variability and dotted line represents the channel variability.

3. Speaker Verification systems

Figure 2 shows the block diagram of the speaker verification system used in this work. SI model refers to speaker independent model. This model is obtained from a large amount of speech data from different speakers. SD model refers to speaker dependent model. It is obtained by maximum-a-posteriori (MAP) adaptation the SI model. Only the means of the Gaussian components were adapted as follows.

$$m_i^{sd} = \alpha * m_i^{si} + (1 - \alpha) * m_i^X,$$

where m_i^{sd} is the i^{th} component of the adapted mean of the speaker model, m_i^{si} is the corresponding mean of the SI model, m_i^X is the mean estimated from the training data, and α is the adaptation factor. During testing, the log-likelihood of the utterance - also referred as score - is calculated with respect to the SI (**LL**SI) model and the SD (**LL**SD) model of the claimed speaker. The difference between SD score and SI score is compared with a threshold to validate the identity of the claimed speaker.

3.1. GMM-based system (baseline system)

Gaussian mixture model is a well known statistical technique for modeling the data. It has been shown that an arbitrary distribution can be modeled using sufficiently large number of mixtures. The log-likelihood of the d-dimensional data $X = \{x_t\}$ for a given GMM Λ is estimated as follows,

$$\log p(X/\Lambda) = \sum_t \log p(x_t/\Lambda)$$

where x_t is a data vector at time t, and

$$p(x_t/\Lambda) = \sum_i w_i * p(x_t/\Lambda_i)$$

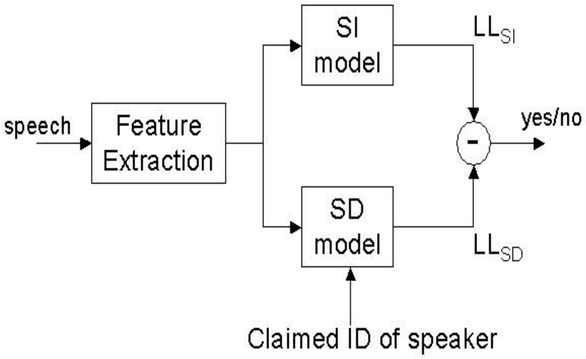

Figure 2: Block digram of the speaker verification system used in this work

where

$$p(x_t/\Lambda_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{(x_t - m_i) * \Sigma_i^{-1} * (x_t - m_i)\right\},$$

and Λ_i is the ith component of the mixture characterized by mean, m_i; covariance, Σ_i; and weight, w_i.

GMM has been shown to be very effective in speaker recognition [5]. In this work, 256 component GMM was used. In SI model training, these components were fitted to the data as follows. First the data were divided into 256 bins using LBG method [9]. Then the means and covariances were re-estimated using Expectation-Maximization (EM) [10] algorithm.

As mentioned before, SD models were derived from SI models using MAP adaptation and only means of the SI model were adapted. During testing, 5 top scoring components were recorded for each frame using SI model. These components were used to estimate the likelihood of the data with respect to the SI and SD models.

3.2. Phone-based system

The choice of phone categories was based on two reasons. First it was observed that the phones within each broad category contain similar amount of speaker and channel variability [8]. Second, the broad categories can be recognized more accurately (65%-75% accuracy) than the phones (45%-55%) using a simple speech recognition system.

In this work, we used 4 broad phonetic categories- vowels+diphthongs, fricatives, glides+nasals, and silence+stops - in the speaker verification system. Each phone category was modeled as hidden Markov model (HMM) using HTK toolkit [11]. HMM is widely used for modeling phonetic units in continuous speech recognition [12, 13]. It is an extension of Markov model technique. Markov models use states and inter-state transitions to model a sequence of observations where the transition between states is modeled using the transition probability matrix. In each state, the distribution of features is modeled using a mixture of Gaussians. Markov models assume that the sequence of states is known and only the probability of observations within each states is trained. HMMs do not make this assumption about the sequence of states and during training the sequence of states needs to be estimated too.

HMMs are described using a triplet $\lambda = (A, B, \pi)$ where $A = \{a_{ij}\}$ is the state-transition probability matrix, $B = \{b_j(k)\}$ is the probability distribution within each state, $\pi = \{\pi_i\}$ is the initial state distribution, i and j are a state indices,

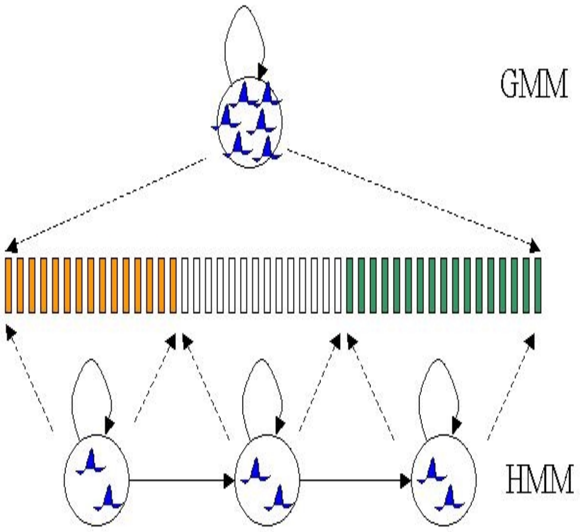

Figure 3: Comparison between GMM and HMM. Circles represent states which contain mixture of Gaussians. Small hats represent Gaussians. Loop for a state is the probability of remaining in that state and line between states is the transition probability between states.

and k is number of components per state. In this work, we have used 1-state, multi-component HMM, i.e., $i = j = 1$ and $k = 1, 2, .., N$ where N is the number of Gaussian components in a state. The probability of the d-dimensional data vector, x_t with respect to a mixture in a state is given by

$$p\,(x_t/b_j) = \sum_k c_k * p\,(x_t/b_j\,(k))\,,$$

where $b_j\,(k)$ is defined similar to Λ_i and c_k is similar to w_i from the GMM equation set.

The phone labels for speaker independent model training data were obtained from the Center for Language and Speech Processing at Johns Hopkins University. Using these labels, HMMs for different categories were initialized by Forward-Backward algorithm [13]. The initialized models were re-estimated using Baum-Welsh algorithm to maximize the likelihood of the data with respect to each category. In the last step, all the phone models were modified simultaneously to maximize the likelihood of the whole utterances using embedded re-estimation algorithm [14].

The training data for the speaker models were labeled using the SI model, Forward-Backward algorithm and Viterbi decoding. For decoding, a simple grammar was used – any category can follow any other category. Using the labels, phone models for each speaker were derived from the SI phone models. During testing, the utterance was labeled using the SI model. The likelihood of the data with respect to the categories was estimated separately. These likelihoods were merged depending upon the type of the experiment.

3.3. Comparison of the two systems

The components of the GMM can be related to a sub-phone class. However, there are significant differences between sub-phone HMM and Gaussian component. Firstly GMM assumes that the consecutive frames are independent. This means that it does not model any temporal information. It does not model the duration of the units either. HMM models the sequence information across states but within a state the observation are assumed to be independent (Note that GMM can also be considered as a 1-state HMM (see Figure 3).). The recognition procedure is very simple for GMMs as the recognition can be performed using only the given frame. For HMMs, the decision about any frame is based on the likelihood whole utterance and Viterbi algorithm used in recognition is computationally expensive.

4. Experimental Setup

In these experiments, NIST 1996 speaker verification development data was used for training the SI model. The data contains conversations from Switchboard-1 corpus. For training, we have used 2 minutes of data from 40 male and 40 female speakers.

For the SD model training and testing, we used 1999 and 2000 NIST speaker verification evaluation data. This contains conversations from Switchboard-2 corpus. The 1999 evaluation data consists of 539 speakers (230 males + 309 females) and 2000 evaluation data consists of 1003 speakers (457 males + 546 females). In both cases, approximately 2 minutes of data was used for training the speaker models. For testing, we used 37620 trials (all the durations) from 1999 evaluation data and the 47797 trials of 15-45 s duration from 2000 evaluation data for the experiments. We analyzed the results under two conditions: 1) when the same handset type is used for training and testing (referred as TELOUT condition) and 2) when different handset types are used for training and testing (referred as HSTOUT condition).

Twenty three Mel frequency cepstral coefficients (MFCCs) were used as features in our experiments. They were estimated using 25 ms speech segment. The adjacent segments were overlapped by 15 ms. As the utterances were recorded over telephone line, 19 filter-banks in the range of 300 Hz - 3300 Hz were used to compute the spectrum. The logarithm of spectrum was computed and it was projected on 19 discrete cosines to obtain MFCCs. First cepstral coefficient (C0) was ignored as it is shown be sensitive to the channel variability. The long term mean over the utterance was removed from each trajectory too, as it is also shown to be sensitive to the channel distortions. Finally delta coefficients were calculated using these features over a 5 frame window and were appended to MFCCs resulting in a 36 dimensional feature vector. We empirically observed that variance normalization of each cepstral coefficient gives best results for the phone based system and whitening [15] of the cepstral features gives the best results for the GMM-based system. Hence system specific feature normalization technique was used.

5. Results

The results are presented in terms of ROC curves and equal error rate (EER) points. This statistics is estimated as follows. During testing, each file was compared with 11 punitive speakers. For each file, the identity of the target speaker was provided by NIST after the evaluation. Using this information the scores for target and imposter speakers were segregated. These scores were sorted and the each score value was used as a threshold. At each step, the false acceptance F_A and the false rejection F_R probabilities were calculated. The resulting curve of F_A verses F_R is called an ROC curve. The EER point on this curve represents $F_A = F_R$ (denoted by a cross on ROC curve). Note

Table 1: Comparison of equal error rate (EER) using different broad phone categories. 1=silence+stops, 2=glides+nasals, 3=vowels+diphthongs and 4=fricatives. Note that the combination (3 4) with 128 component mixture performs comparable to the complete system with 256 component mixtures.

Broad Phone Category	%EER		Gaussian Components
	TELOUT Condition	HSTOUT Condition	
1	22.8	30.9	64
2	18.4	26.8	
3	17.2	24.3	
4	16.7	23.7	
1 2	18.0	25.9	128
1 3	17.4	25.7	
1 4	17.9	26.0	
2 3	15.5	23.5	
2 4	15.2	23.1	
3 4	**14.0**	**21.3**	
1 2 3	15.5	23.4	192
1 2 4	15.7	23.6	
1 3 4	15.4	22.7	
2 3 4	14.0	22.3	
1 2 3 4	**14.3**	**21.9**	256

that EER weights both the probabilities equally. NIST penalizes F_A more than F_R and the corresponding optimal point is represented by a circle on ROC curve.

5.1. Verification using broad phonetic categories

Table 1 shows the speaker verification results using the four broad categories. It shows that fricatives performs better than glides+nasals in both TELOUT and HSTOUT conditions. Fricatives also perform comparable to vowels+diphthongs. This surprising result was attributed to under-fitting of the 1 state, 64 component HMM for the vowels+diphthongs category, which has approximately twice the data as fricatives. This conclusion was verified by increasing the components of HMM from 64 to 128. With new models, it was observed that vowels+diphthongs outperformed the fricatives category. Note that these results are in agreement with the results based on ANOVA (see section 2).

Table 1 also shows the results obtained by combining different phonetic categories. The combination of silence+stops with other categories does not change their individual performance. Combination of glide+nasals and vowels+diphthongs gives similar results as combination glide+nasals and fricatives. The best results are obtained using vowels, diphthongs and fricatives categories. Adding silence+stops to this degrades the performance and adding glides+nasals to this combination does not improve the performance further. Note that the combination of vowels+diphthongs and fricatives with 128 components performs comparable to the complete system with 256 components.

5.2. Comparison of the SPVER systems

For these experiments we used gender dependent SI models. The phone-based system was modified based on the results of the previous experiment. Vowels and diphthongs were modeled using 128 component HMM. The remaining categories were modeled using 1-state, 64 component HMMs. While testing, the likelihood for silence and stops category was excluded from the calculation of likelihood of the utterance.

Figure 4 shows the performance of the phone-based system and GMM-based system on the NIST speaker verification evaluation tasks. It shows that the phone-based system outperforms the GMM-based system in both TELOUT and HSTOUT conditions. For 2000 evaluation task, EER for phone-based system was 11.8 % and 17.9 % and for baseline system was 12.6 % and 22.4 % for TELOUT and HSTOUT conditions respectively. For 1999 evaluation task, EER for phone-based system was 10.2 % and 18.8 % and for baseline system was 12.5 % and 26.2 % for TELOUT and HSTOUT conditions respectively.

The performance improvement using the phone-based system is attributed to two factors. First, note that both GMM-based and phone-based systems ignore around 30% of the test data. But the fraction of frames ignored within each category is different in two systems. The GMM-based system uses energy based speech-silence segmentation and ignores the frames with low energy. Therefore GMM-based system would ignore unvoiced fricatives but would use voiced stops during testing. The phone-based system performs the speech-silence segmentation by ignoring the silence+stops category. Therefore it uses both the voiced and unvoiced fricatives for testing and ignores both voiced and unvoiced stops. Our experiments have shown that ignoring the unvoiced fricatives degrades the performance of the phone-based system.

Second factor is about the modeling. It has been shown that there is an overlap between glides and vowels in the feature space. In GMM-based system, the Gaussian components in the common space are shared by both the categories. The phone-based system, however, would model the common space twice; once for each category. This might be a suboptimal strategy as far as modeling is concerned. But the speaker information represented by two categories might be different and a separate modeling of these categories might be beneficial for the SPVER task.

6. Conclusions

In the previous work, speaker and channel variability in 6 broad phonetic categories - vowels, diphthongs, fricatives, glides, nasals and stops was analyzed using ANOVA. The results showed that vowels, diphthongs, nasals, and fricatives contain the most useful speaker variability. The stops were observed to vary the least across different speaker variability.

Based on these results, we designed a speaker verification system that uses 4 broad phonetic categories: vowels+diphthongs, glides+nasals, fricatives, and silence+stops. We compared the performance of each phonetic category using NIST speaker verification evaluation data. The results show that vowels and diphthongs are the most useful categories for

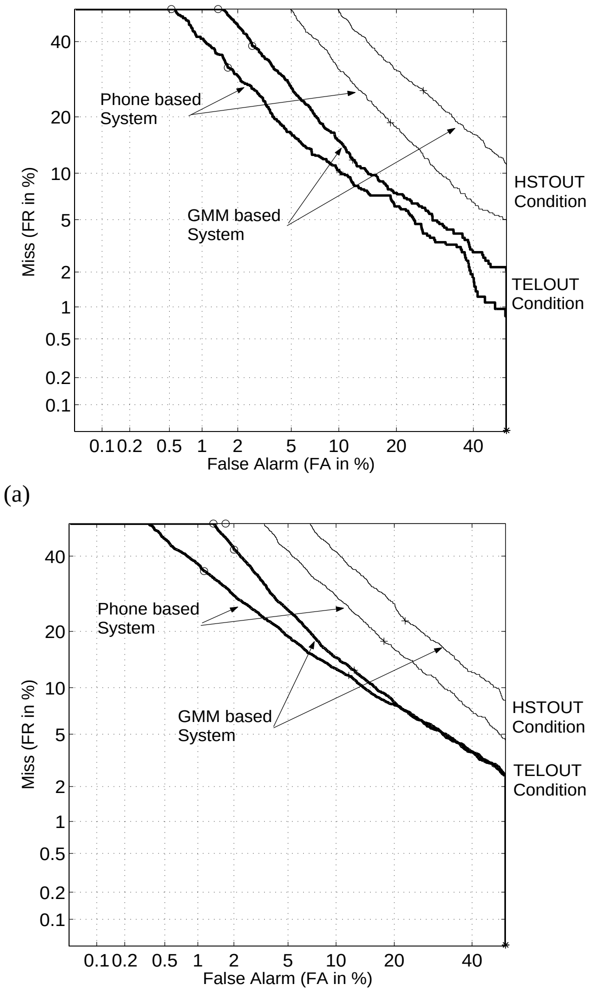

Figure 4: (a) Eval 99 performance and (b) Eval 00 performance. Cross indicates EER points where FA=FR. Note that phone-based system outperforms the GMM-based system in both TELOUT and HSTOUT conditions.

speaker verification and silence and stops are the least important categories. These results confirm our conclusions from the analysis of phone specific speaker and channel variability in speech. They also agree with the previous phone-based speaker verification results [1, 2, 16, 3].

The speaker verification system was modified based on these results. The category of vowels and diphthongs was modeled using 1 state, 128 component HMM and 1 state, 64 component HMM was used for the other categories. The silence and stops category was ignored during testing phase. This system was compared with the 256 component GMM-based system on the 1999 and 2000 NIST speaker verification evaluation tasks. Note that both the systems used same number of components in testing. Both systems also ignored approximately 30 % of the data - GMM based system uses energy-based threshold and phone-based system ignores stops+silence category. The results showed that the phone-based system outperforms the GMM based system in both TELOUT and HSTOUT conditions.

7. Acknowledgements

Authors thank Shankar Kumar and Bill Byrne from CSLP, Johns Hopkins University for providing the phone labels for the NIST 1996 speaker verification development data. The research was supported by NSA/DoD under MDA904-00-C-2089, NSF under IRI-9712579 and by an industrial grant from Intel Corporation.

8. References

[1] S. K. Gupta and Michael Savic, "Text-independent speaker verification based on broad phonetic segmentation of speech," *Digital Signal Processing*, vol. 2, pp. 181–202, 1992.

[2] E. S. Paris and M. J. Carey, "Discriminative phonemes for speaker identification," in *ICSLP*, 1994, pp. 1843–1846.

[3] J. Koolwaaij and J. de Veth, "The use of broad phonetic class models in speaker recognition," in *ICSLP*, 1998.

[4] Frederick Weber, Barbara Peskin, Michael Newman, Andres Corrada-Emmanuel and Larry Gillick, "Speaker recognition on single- and multispeaker data," *Digital Signal Processing*, vol. 10, pp. 75–92, 2000.

[5] T. F. Quatieri D. A. Reynolds and R. B. Dunn, "Speaker verification using adapted mixture models," *Digital Signal Processing*, vol. 10, pp. 181–202, 2000.

[6] D.A. Reynolds, "Speaker identification and verification using gaussian mixture models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[7] Sachin Kajarekar, Naren Malayath and Hynek Hermansky, "Analysis of sources of variability in speech," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 343–346.

[8] Sachin Kajarekar, Naren Malayath and Hynek Hermansky, "Analysis of speaker and channel variability in speech," in *Proc. of ASRU*, Colorado, 1999.

[9] A. Buzo Y. Linde and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Comm.*, vol. COM-28, pp. 84–95, Jan. 1980.

[10] N.M. Laird A.P. Dempster and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[11] S. Young et. al., *The HTK Book*, Cambridge University, 1997.

[12] F. Jelinek, "Continuous speech recognition by statistical methods," *IEEE Proceedings*, vol. 64, pp. 532–556, Apr. 1976.

[13] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[14] Steve Young et. al., *The htk Book*, Entropic, first (revised) edition, 1999.

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 2nd ed., 1990.

[16] L. Rodriguez-Lunares and C. Garcia-Mateo, "On the use of acoutsic segmentation in speaker identification," in *Eurospeech*, 1997.