A CONTINUOUS HMM TEXT-INDEPENDENT SPEAKER RECOGNITION SYSTEM BASED ON VOWEL SPOTTING

Nikos Fakotakis, Kallirroi Georgila*, Anastasios Tsopanoglou***
* Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 26110 Rion, Patras, Greece
Tel: +30 61 997336, Fax:+30 61 991855, e-mail: fakotaki@wcl.ee.upatras.gr, rgeorgil@wcl.ee.upatras.gr

** KNOWLEDGE S.A., Human Machine Communication Dept.,
N.E.O. Patron-Athinon 37, 264 41 Patras, Greece
Tel: +30 61 452820, Fax:+30 61 453819, e-mail:KNOWLEDGE@Patra.hol.gr

ABSTRACT

This paper presents a text-independent speaker recognition system based on vowel spotting and Continuous Mixture Hidden Markov Models. The same modeling technique is applied both to vowel spotting and speaker identification/verification procedures. The system is evaluated on two speech databases, TIMIT and NTIMIT, resulting in high accuracy rates. Closed-set identification accuracy on TIMIT and NTIMIT databases is 98.09% and 59.32%, respectively. Concerning the verification experiments, accuracy of 98.28% for TIMIT, and 83.04% for NTIMIT databases is obtained. The nearly real time response of the classification procedure, the low memory requirements and the small amount of training and testing data are some of the additional advantages of the proposed speaker recognition system.

1. INTRODUCTION

Depending upon the application, the general area of speaker recognition is divided into two specific tasks: identification and verification. In speaker identification, a speaker is identified among several speakers of known voices. This is also referred to as closed-set speaker identification. In verification, the goal is to decide whether a speaker is the person he/she claims to be. The speaker identification and verification may be either text-dependent or text-independent. The former approach assumes that the speakers utter a specific phrase, a sequence of words or a predefined password while the latter verifies the speaker's identity regardless of the content of the utterance.

The performance of the existing speaker recognition systems seems to be very high in comparison to other applications concerning the speech processing scientific area. Nevertheless, a great amount of work is still to be done to approximate the optimum target of perfection, especially when the system works in real circumstances, in a noisy environment and when the training and testing data are limited and completely non-correlated.

In this work, an efficient text-independent speaker recognition system is described. The classification of the speakers is carried out by using only the vowel segments of the input speech signal. Assuming that adequate information about the speaker's characteristics is contained within the vowel segments of the speech signal the method seems to be very efficient and promising. The idea of using the vowel parts of the speech signal was initially introduced and developed in 1986 [1] and improved later [2],[3]. The technique proved to be a very effective solution to the speaker classification problem and the initial obtained results were very promising and encouraging.

In the present system, the basic idea of vowel spotting is still retained, but several essential changes have been made to improve the accuracy of the system and to test the way in which the vowel spotting procedure can be combined with the well known Hidden Markov Modeling techniques. The use of Gaussian mixture speaker models has demonstrated high text-independent identification and verification accuracy for short test utterances for both clean and telephone quality speech [4].

The system is evaluated on two widely available speech databases: TIMIT and NTIMIT. The use of the TIMIT database aims at examining the performance of a text-independent speaker recognition system under near-ideal conditions using a large population, while the NTIMIT is used to gauge the error due to the presence of noise in the telephone network, for the same large population experiment.

The rest of the paper is organized as follows. A detailed presentation of the recognition system is given within Section 2, which in its turn is separated into three subsections: the feature extraction, the vowel spotter and the speaker classifier. The system training procedure is described in Section 3 and the speech databases used together with the experimental results, are demonstrated within Section 4. Finally, a short summary and some conclusions are submitted in Section 5, which is followed by references in Section 6.

2. SYSTEM DESCRIPTION

The speaker recognition system consists of three functional modules, shown in Figure 1. The feature extraction, the vowel spotting, and the speaker classification and threshold comparison module.

In the first module, the parameter vectors, that contain information about the temporal and frequency characteristics of the speech signal, are extracted. The second module is responsible for the location of the vowel segments within the input speech signal and the corresponding parameter vectors. According to this approach, the number of used parameter vectors is significantly reduced and both the training and testing procedures are accelerated. The speaker classifier incorporates the output of the vowel spotter and computes the probabilities of each speaker for the specific utterance. Finally, in the case of speaker verification, the estimated probability of the claimed speaker is compared to the speaker dependent threshold and the system decides if the claim of the speaker is valid.

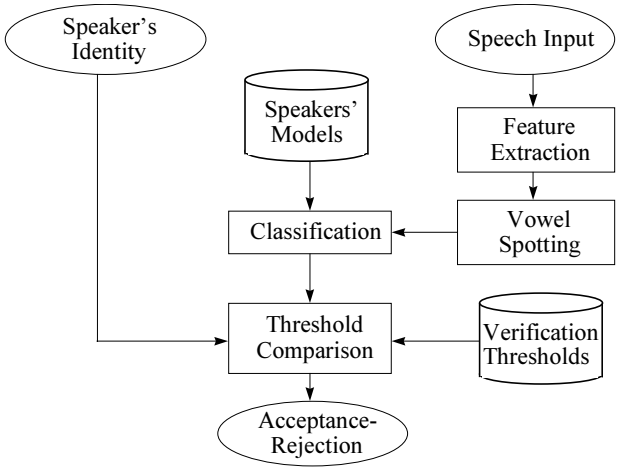

Figure 1. *Block diagram of the text-independent speaker recognition system.*

2.1. Feature Extraction

The speech signal is sampled at 16KHz and digitized at 16 bit resolution. The signal is segmented into frames using a Hamming window of 20ms length with a step size of 10ms resulting in an overlap of 10ms. Each frame is preemphasized using a constant factor of a=0.95. At the next step, 19 mel-scale cepstral coefficients (MFCC), 19 first order differential MFCC and 1 normalized energy coefficient are calculated for each temporal frame. The resulting 39-dimension vector is applied to the vowel spotter.

2.2. Vowel Spotting

Concerning the vowel spotting module, a separation of the phonemes of the speech database into two categories is carried out. The produced phoneme sets are: the vowels and the non-vowels phonemes. Each category is modeled by a 3 state, left-to-right Hidden Markov Model. The states describe the beginning, the middle and the end of each vowel phoneme. Seven Gaussian mixtures per state, corresponding to the number of phoneme clusters, are used to describe the observations classified to each state. The segmental k-means algorithm is applied to the speech segments, corresponding to each HMM state of each model to provide the number of mixtures and compute the mean and the variance vectors of each mixture [5]. If the number of segments within a state is not large enough, the number of mixtures is reduced for that state. The Viterbi algorithm is used for the computation of the model parameters and the Level Building algorithm classifies the phonemes into the two above mentioned categories during the vowel spotting procedure [6].

2.3. Speaker Classification

In the speakers' models database, each speaker is described by a 3 state ergodic Gaussian mixture HMM, which has 5 mixtures per state. The segmental k-means is used for the computation of the mean and the variance vectors of each mixture and the state/segment duration and normalized energy distributions have been incorporated. The model parameters are computed by the Viterbi algorithm and the Level Building recognition algorithm is applied to find out the speaker with whom the unknown observation sequence of the vowels matches better.

For speaker verification, the derived probability from the speaker classification module is compared to the threshold of the speaker requested for verification. The result of the comparison is taken into consideration in order to decide if the claim of the speaker has to be accepted or rejected.

3. TRAINING PROCEDURE

The training procedure of the system, shown in Figure 2 involves three modules: the vowel spotter, the speaker classifier and the verification threshold estimation module.

The vowel spotting and the speaker classification training procedures are performed sequentially. The vowel spotting module is speaker independent and is trained using a large amount of data that are non-correlated to the ones used for training the speaker classification module or testing the system. The two models (vowels, non-vowels) are the output of the vowel spotting training procedure. These models are used to locate the vowel parts of the speech signal that will form the speakers' feature sets. For training the speakers' models a different database is used, from those used to train the vowel spotter and to test the system. The parameters of the

speakers' models database are calculated during the speaker classification training module.

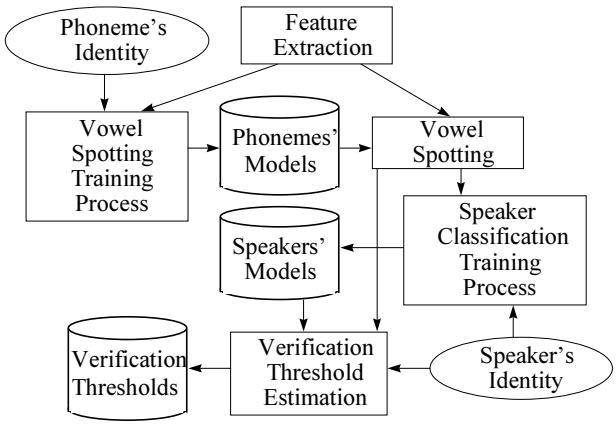

Figure 2.*Training procedure of the speaker recognition system.*

The verification procedure for each speaker is based on a decision threshold which is estimated during the training phase. The decision threshold that is used, is based on the Minimum-Error threshold theory and is derived from the inter- and intra-speaker distance distributions, at the point where the probability of the mean verification error is minimum [1],[2]. The mean verification error is defined as the mean value of the false rejection and false acceptance errors. In the present system, the derived probabilities are transformed into distance-like measures and are used to define the verification thresholds for each speaker involved in the system.

The mean value μ_{1i}, and the standard deviation σ_{1i}, of the intra-speaker distribution for the ith speaker are calculated by:

$$\mu_{1i} = \frac{1}{N}\sum_{k=1}^{N}\log(p_{ik}) \qquad and \qquad \sigma_{1i}^2 = \frac{1}{N}\sum_{k=1}^{N}\left[\log(p_{ik}) - \mu_{1i}\right]^2,$$

where N is the number of training observation sequences for the ith speaker concerning his/her utterances and p_{ik} corresponds to the produced probability when the model λ_i of the ith speaker is used, for the kth utterance.

The mean value μ_{2i}, and the standard deviation σ_{2i}, of the inter-speaker probability distribution for the ith speaker are calculated by:

$$\mu_{2i} = \frac{1}{M}\sum_{j=1}^{M}\log(p_{ij}) \qquad and \qquad \sigma_{2i}^2 = \frac{1}{M}\sum_{j=1}^{M}\left[\log(p_{ij}) - \mu_{2i}\right]^2,$$

where M is the number of probabilities of the ith speaker corresponding to the utterances of the remaining speakers, that is $N_U(N_s\text{-}1)$, where N_U is the number of utterances that each speaker utters, N_s is the number of speakers that are used for training and p_{ij} is the probability of the ith speaker for the jth utterance.

The Minimum-Error threshold of the ith speaker is given by:

$$\theta_i = \frac{\mu_{1i} - \mu_{2i}\sigma_{Ri}^2 - \sigma_{Ri}\sqrt{(\mu_{1i} - \mu_{2i})^2 - 2\sigma_{2i}^2(1-\sigma_{Ri}^2)\ln\sigma_{Ri}}}{1-\sigma_{Ri}^2},$$

where $\sigma_{Ri} = \sigma_{1i} / \sigma_{2i}$ is the ratio of the standard deviations.

4. EXPERIMENTAL RESULTS

4.1. Speech Databases

The system has been trained and tested on two publicly available databases, TIMIT and NTIMIT, the former for clean speech and the latter for telephone quality speech.

The TIMIT database contains 630 speakers (438 male and 192 female) each of them having uttered 10 sentences. Each utterance is a read sentence of approximately 3 seconds. The sentences have been designed to have rich phonetic variability and have been chosen from the 8 major dialect divisions of the United States. The speech signal is recorded through a high quality microphone, in a very quiet environment. For each speaker all the recordings took place in a single session [7].

The NTIMIT database was obtained by playing TIMIT speech signal through an "artificial mouth" installed in front of a carbon-button telephone handset via a telephone test frame designed to approximate the acoustic coupling between the human mouth and the telephone handset. The speech signal was transmitted through a local or long-distance network of a different telephone line for each sentence [8].

In order to train and test both vowel spotter and speaker classifier, the TIMIT and NTIMIT databases have been divided into 4 sets. The first set consists of 15 speakers (10 male and 5 female) from every region, resulting in a population of 120 speakers, each having 10 utterances. Five different utterances of 410 speakers (288 male and 122 female) comprise the second set. The remaining 5 utterances of the 410 speakers used for training the speakers' models, form the third set. Finally, the fourth set consists of 10 utterances from the remaining 100 speakers (70 male and 30 female) used as impostors in the speaker verification process.

The first set of the TIMIT and NTIMIT databases, as defined above, is used for training the vowel spotting procedure. The second set is used to train the speaker classifier. During the classifier's training process the Minimum-Error Threshold is calculated for each speaker. The procedure is implemented for both TIMIT and NTIMIT.

4.2. System Testing

The testing process concerns evaluating the performance of the vowel spotting, the speaker identification and speaker verification procedures.

4.2.1. Vowel Spotting

The second and the third sets of TIMIT and NTIMIT are applied to testing the vowel spotting process, which takes place during the vectors' creation of the speakers and during the speaker classifier's evaluation.

For the TIMIT and NTIMIT speech databases, the percentage of the correctly recognized vowels is 71.51% (with 3.42% false acceptance error rate) and 61.92% (with 6.12% false acceptance error rate) respectively. The above proportion may not be very high but it is proven adequate for the correct function of the system. It should be noted that the proportion of the semivowels and nasals recognized as vowels is 3.39% and the proportion of consonants is only 0.03%, for TIMIT. Tests performed with four models (vowels, semivowels, nasals and the remaining phonemes) show that for spotting not only the vowels but also the semivowels and the nasals in the same cluster, the recognition accuracy reaches 92.16% for TIMIT and 86.15% for NTIMIT.

4.2.2. Speaker Identification

The closed-set identification test is performed upon the third set of both TIMIT and NTIMIT databases. Table 1 demonstrates the results of closed-set identification tests for male and female population and the total recognition accuracy.

	Male (%)	Female (%)	Total (%)
TIMIT	97.78	98.85	98.09
NTIMIT	59.17	59.67	59.32

Table 1. *Closed-set identification results.*

4.2.3. Speaker Verification

The verification tests are performed upon the third set of databases and the 100 speakers, that comprise the fourth set of the databases as impostor speakers. Tables 2 and 3 show the false rejection (FR), the false acceptance (FA) and the mean ((FA+FR)/2) error rates of the verification test for both TIMIT and NTIMIT databases.

TIMIT/ NTIMIT	FR Error (%)	FA Error (%)	Mean Error (%)
Male	1.32/15.76	2.31/19.03	1.81/17.39
Female	1.14/14.59	1.89/17.27	1.51/15.93
Total	1.27/15.41	2.18/18.51	1.72/16.96

Table 2. *Verification results for TIMIT and NTIMIT databases(FR=False Rejection, FA=False Acceptance).*

5. SUMMARY AND CONCLUSIONS

This paper has presented and evaluated a text-independent speaker recognition system based on vowel spotting and Continuous HMMs. The Minimum-Error threshold was used for deciding on accepting or rejecting the claimed speaker. Experimental evaluation of the system's performance was conducted on two publicly available databases, TIMIT for clean speech and NTIMIT for telephone quality speech.

The system is an efficient speaker recognition system with identification accuracy 98.09% and 59.32%, and verification accuracy 98.28% and 83.04% for TIMIT and NTIMIT respectively. A significant fall to the system performance is observed when the system is used in a noisy environment (NTIMIT database) but current work is carried out to improve its accuracy. It should be noted that there has been no correlation between training and testing data throughout all the experiments.

6. REFERENCES

[1] N. Fakotakis: *"A New Method of Automatic Speaker Recognition"*, PhD thesis, University of Patras, 1986.

[2] N. Fakotakis, A. Tsopanoglou and G. Kokkinakis: *"A Text-Independent Speaker Recognition System Based on Vowel Spotting"*, Speech Communication Journal, Vol. 12, No. 1, pp. 57-68, March 1993.

[3] N. Fakotakis and J. Sirigos: *"A High Performance Text-Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets"*, Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 661-664, May 7-10, 1996, Atlanta USA.

[4] D. A. Reynolds and R. C. Rose: *"Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models"*, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83, Jan. 1995.

[5] L. R. Rabiner, J. G. Wilpon and B. H. Juang: *"A segmental k-means training procedure for connected word recognition"*, AT&T Technical Journal, vol. 65, pp. 21-31, May/June 1986.

[6] L. R. Rabiner, J. G. Wilpon and F. K. Soong: *"High Performance Connected Digit Recognition Using Hidden Markov Models"*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 8, pp. 1214-1225, August 1989.

[7] W. Fischer, V. Zue, J. Bernstein and D. Pallet: *"An Acoustic-Phonetic Database"*, JASA, Suppl. A. Vol. 81 (592) 1986.

[8] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz: *"NTIMIT: A phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database"*, Proc. Internat. Conf. Acoust. Speech Signal Process., April 1990, pp. 109-112, New Mexico, USA.