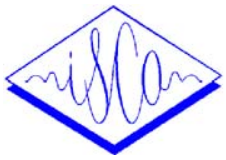

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HIGH PERFORMANCE TEXT-INDEPENDENT SPEAKER RECOGNITION SYSTEM BASED ON VOICED/UNVOICED SEGMENTATION AND MULTIPLE NEURAL NETS

Nikos Fakotakis, John Sirigos and George Kokkinakis
Wire Communications Laboratory
University of Patras, 26500 Greece
fakotaki@wcl.ee.upatras.gr

10.21437/Eurospeech.1999-239

ABSTRACT

This paper presents a text-independent speaker recognition system based on the voiced segments of the speech signal. The proposed system uses feedforward MLP classification with only a limited amount of training and testing data and gives a comparatively high accuracy. The techniques employed are: the Rasta-PLP speech analysis for parameter estimation, a feedforward MLP for voiced/unvoiced segmentation and a large number (equal to the number of speakers) of simple MLPs for the classification procedure. The system has been trained and tested using TIMIT and NTIMIT databases. The verification experiments presented a high accuracy rate: above 99% for clean speech (TIMIT) and 74.7%, for noisy speech (NTIMIT). Additional experiments were performed comparing the proposed approach of using voiced segments with only vowels and all phonetic categories with results favorable to the use of voiced segments.

1. INTRODUCTION

The task of a speaker recognition system is either to identify an unknown speaker among several speakers of known speech characteristics or to verify whether a speaker is the person he claims to be. Speaker recognition may be text-dependent or text-independent. Text-dependent speaker recognition systems require that the speaker utters a specific phrase or a given password. Text-independent speaker identification systems identify the speaker regardless of his utterance. Text-independent recognition systems are more versatile but their accuracy is considerably lower than that of comparable text-dependent systems. The identification process can be closed set or open set. Closed-set speaker identification refers to the case where the speaker is known a priori to be a member of a set of N speakers. Open-set speaker identification includes the possibility that the speaker does not belong to the set of N speakers.

The most popular modern techniques for speaker recognition are based on Hidden Markov Models (HMMs) [1], Artificial Neural Networks (ANNs) [2] and combinations of both which lead to Hybrid approaches [3].

In this paper we present a text-independent ASR system which is suitable both for identification and verification purposes. This system separates the voiced part of the speech signal, using a voiced/unvoiced decision module, to form feature vectors and feedforward MLP classifiers for speaker models. The idea of using only the voiced part of the speech signal is based on the fact that voiced speech segments contain the most significant speaker identification information as opposed to other speech segments [4]. The benefits of using voiced segments are presented in detail in section 3, where comparable results of experiments using various phonetic categories are given.

In section 2, a detailed description of the recognition system is presented. The experimental results are given in section 3. The system has been evaluated on both identification and verification experiments, using clean and telephone speech. The paper concludes with some remarks in section 4.

2. SYSTEM DESCRIPTION

The block diagram of the presented system is shown in Fig. 1. Initially feature vectors are extracted from the input signal through signal processing. Then, the system separates the

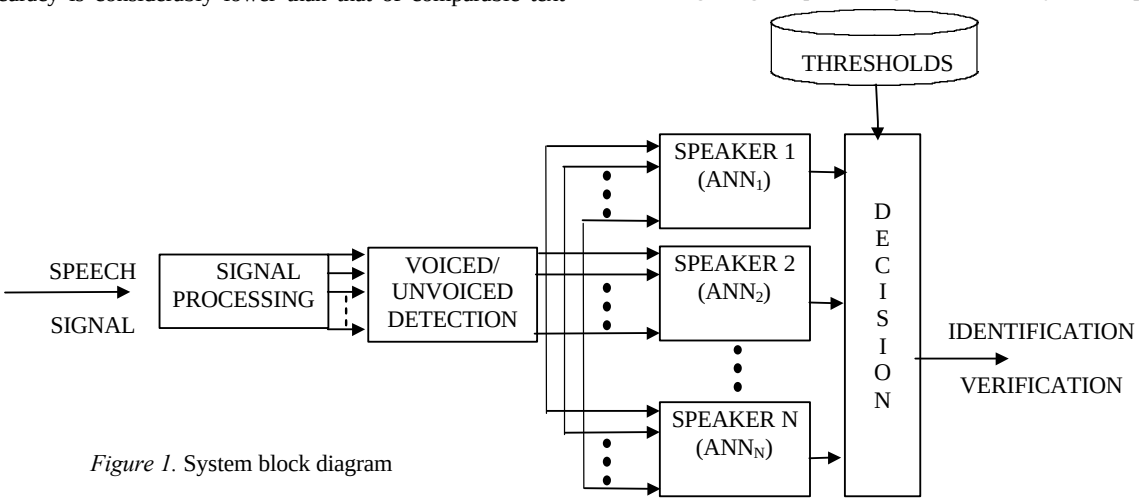

Figure 1. System block diagram

voiced part of the speech signal, using a voiced/unvoiced decision module. The last module performs speaker classification, providing the final decision for identification or verification. Below we describe analytically each one of the modules.

2.1 Signal Processing

Speech input, sampled at 16 kHz and digitized at 16-bit resolution, is provided to the signal processing module of the system. The speech signal is segmented into frames of 30 ms duration with 15 ms overlapping. After the application of a Hamming window each frame is analyzed using the Rasta PLP speech analysis technique in order to obtain the characteristic parameters of the signal. An autoregressive 12^{th} order all-pole model approximates the resulting spectrum of the speech signal from which cepstral coefficients are computed.

Thus, at every q^{th} frame a feature vector of $\{c_i(q)\}$;i=1,2,…,P, (P=12) cepstral coefficients is calculated.

2.2 Voiced/Unvoiced Segmentation

The voiced speech is located and identified as relevant speech events with a speaker independent voiced/ unvoiced (V/U) segmentation module which is based on a 3-layer feedforward MLP classifier and on a number of heuristic rules.

The architecture of the ANN has been experimentally chosen to be 12x5x4x1 (9 hidden nodes in two layers). Results concerning the size of this network are shown in section 3. At each frame the 12- dimensional parameter vector extracted from the signal processing stage, feeds the input of the network. Two hidden layers follow with 5 and 4 inputs respectively and finally one output unit determines whether the input vector corresponds to a voiced or an unvoiced frame.

The ANN classifier for the V/U module is trained to be speaker- independent using labeled training data from a large number of speakers, as described in the next section. The feature vectors corresponding to voiced frames are labeled as “one” while the feature vectors from the unvoiced ones are labeled as “zero”. In order to polarize the training procedure and thus minimize the false acceptance error rate, which is more important, we used more data from the unvoiced patterns (75% of the total) than of the voiced ones. This is necessary because the ANN-models of the speakers are trained only with voiced segments. Thus, unvoiced input would produce large errors if the training samples were not polarized.

Initially, all input frames are marked with “one” or “zero” corresponding to a raw decision “voiced, unvoiced”, respectively. Following this raw decision, a procedure based on experimentally derived heuristic rules is applied. The rules take into account the distance between successive frames, the duration of the voiced segments and their amplitude. They assume that the cost of false-rejecting a voiced frame is much less than the cost of false-accepting an unvoiced one. Thus on the output y(q) of the MLP, which is bound between -1.0 and +1.0, the following six-step process is applied to each q^{th} candidate voiced frame:

Step 1: Initialize the internal parameters R_i ($R_1=R_2=R_3=R_4=0$).
Step 2: If y(q)-y(q-5)>0.23, then $R_1=0.15$.
Step 3: If y(q+5)-y(q)>0.37, then $R_2=0.15$.
Step 4: If $\frac{1}{10}\sum_{l=q-5}^{q+5} y(l) > 0.29$, then $R_3=0.4$.
Step 5: If (y(q-2)>0.2) & (y(q-1)>0.18) & (y(q)>0.21) & (y(q+1)>0.24), then $R_4=0.4$.
Step 6: If $\sum_{i=1}^{4} R_i > 0.5$, then the q^{th} frame is considered voiced.

2.3 Speaker Classification

A basic problem in ASR is that by using only one large network, the training time increases exponentially with the number of categories. There are several possible ways to partition a large classification task. In this paper, the speaker classification stage of the system is implemented by modeling each speaker with an individual feedforward MLP network with an experimentally chosen architecture of 12x4x2x1.

The voiced/unvoiced segmentation process filters the speech signal allowing only the voiced frames, represented as 12-dimensional parameter vectors, to feed the MLP classifiers which correspond to the set of speakers known to the system.

For speaker verification the selected feature vectors are applied to the speaker model of the speaker to be verified. A measure is then calculated for the likelihood that this speaker generated the feature vectors. If this measure exceeds a given decision threshold, then the speaker is verified, otherwise he/she is rejected.

The decision threshold used by the proposed system is based on the Minimum-Error (ME) threshold [5], derived from the inter-speaker distance distribution (distances of a centroid from its own training patterns) and intra-speaker distance distribution (distances of a centroid from the training patterns of all other clusters), at the point where the probability of the average error rate (false acceptance and false rejection) is minimum. The analytical expression for the ME-threshold for the i^{th} speaker as a function of the mean values (μ_{1i}, μ_{2i}) and standard deviations (σ_{1i}, σ_{2i}) of the inter- and intra-speaker distance distributions, is given by

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

where ln(.) is the natural logarithm function and $\sigma_{Ri} = \sigma_{1i}/\sigma_{2i}$ is the ratio of the standard deviations. The mean values and the standard deviations of the inter- and intra-speaker distance distributions are calculated as follows:

$$\mu_{1i} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij} \quad (2)$$

$$\mu_{2i} = \sum_{m=1}^{M}\sum_{j=1}^{N_m} |y_{mj} - y_{ij}| \Big/ \sum_{m=1}^{M} N_m \quad (3)$$

$$\sigma_{1i}^2 = \frac{1}{N_i}\sum_{j=1}^{N_i} \{y_{ij} - \mu_{1i}\}^2 \quad (4)$$

$$\sigma_{2i}^2 = \sum_{m=1}^{M}\sum_{j=1}^{N_m} \{|y_{mj} - y_{ij}| - \mu_{2i}\} \Big/ \sum_{m=1}^{M} N_m \ , \quad (5)$$

where y_{ij} is the output of the network corresponding to the j^{th} input sample of the i^{th} speaker. M is the total number of speakers included in the reference set and N_m is the number of samples per speaker.

The MLP classifiers for the individual speaker models are trained using as supervised training procedure a fast version of the back propagation algorithm [6]. Each model is trained using the training data of all speakers in the population. The feature vectors of the same speaker are labeled as “one” and that of the remaining speakers (inhibitory vectors) as “zero”. Thus an MLP trained for each speaker is taken.

Problems appear for large speaker populations, where the large majority of the training vectors have “zero” labels and the classifier tends to learn that everything is “zero”. This problem

is faced by compressing the data representing the inhibitory vectors using a codebook for each speaker and a VQ compression procedure. Thus, the data for each inhibitory speaker are compressed to a small number of vectors prior to constructing the training set for a given speaker. A modified k-means algorithm was used to create the VQ codebook. The codebook size has been experimentally chosen to 16.

In order to adapt a new speaker to the system reference set, a new MLP is added to the network and is trained with the procedure described above. Thus it is not necessary to retrain all the other MLPs in the set, especially when the number of speakers in the reference set is large. Using the weights that already exist, the training time is significantly reduced.

In order to reduce the stored data we keep a record of only the parameter vectors extracted from the voiced parts of the reference data. Thus, from every i^{th} speaker we use as training data about 7,200 bytes (12 parameters x 300 voiced frames x 2 bytes) included in the reference set together with the individual ANN models and the individual decision thresholds J_i. Since all training data are not recorded at the same time, information about the recording date is also kept.

3. EXPERIMENTAL RESULTS

For our experiments we used the TIMIT and NTIMIT databases. The TIMIT database contains 630 speakers (438 male and 192 female) each of them having uttered 10 sentences. Each utterance has approximately a 3 second duration. The sentences are designed to have a rich phonetic variability and have been chosen from 8 major dialect regions of the United States. The NTIMIT database was obtained by playing TIMIT speech signals through an "artificial mouth" installed in front of a carbon-button telephone handset. The speech signal was transmitted through a local or long-distance network of a different telephone line for each sentence.

The MLP model used for the V/U detection module (12x5x4x1), is trained using part of the labeled data from the two speech databases. This process always precedes the overall training procedure for the speaker recognition system. The training data for these two procedures are not correlated. The V/U detector is trained only once; it is speaker independent and does not require further retraining.

To set up the parameters of the V/U decision module, a number of experiments were performed. The following two sets of experiments were performed to evaluate the V/U detection module and the results obtained are shown in Table I. The average error rate (AER), is calculated as a function of the falsely accepted (FA) frames and the falsely rejected (FR) frames. The frames are labeled as voiced or unvoiced.

Experiment 1: Training and testing data from the same dialect regions (Training data: 120 speakers selected from all 8 dialect regions - 10 utterances per speaker, Testing data (uncorrelated to training data): 160 speakers selected from all 8 dialect regions - 10 utterances per speaker).

Experiment 2: Training and testing data from different dialect regions (Training data: 120 speakers - 10 utterances per speaker, Testing data: 160 speakers - 10 utterances per speaker. The dialect regions have been randomly selected)

Training Data	*Testing Data*	*Experiment 1 AER(%)*	*Experiment 2 AER(%)*
TIMIT	TIMIT	1.68	2.28
NTIMIT	NTIMIT	4.14	4.44
TIMIT	NTIMIT	12.75	14.21

Table 1: V/U detection module performance.

The V/U segmentation module used in the integrated speaker recognition system has been trained as in the first experiment. 120 speakers were used for training and the remaining 510 speakers were used for training and testing the overall system. This process was repeated twice once for TIMIT (training and testing) and once for NTIMIT (training and testing).

For the verification tests, each test utterance was used both as one true speaker test utterance and as M-1 impostor test utterances (M=200), rotating through all speakers. The false acceptance error rate, FA (falsely accepted trials, over the total number of crosstrials) and the false rejection error rate, FR (falsely rejected trials, over the total number of autotrials) were measured to calculate the average error rate (AER=(FA+FR)/2,%).

Experiments were performed on the TIMIT and NTIMIT databases with 200 speakers (25 speakers, 17 male and 8 female, from each-dialect region), to evaluate the performance of the speaker verification system. The verification error rate as a function of the amount of training data used for training the individual MLP models, was also measured. We used as training data voiced phonemes with variable length from the first five utterances. The voiced phonemes from the remaining five utterances were used as test data and the results obtained are summarized in Table 2. Experiments with different phonetic categories for training and testing data are also shown in this table. We have trained the speaker models by first using all the phonetic units, then only vowels and finally only the voiced parts of the speech signal and we compared the results. The use of the voiced frames outperformed the other phonetic categories.

Training Data per speaker	TIMIT Database			NTIMIT Database		
	AER(%)			AER(%)		
Phonemes	*All*	*Vowel phonemes*	*Voiced segments*	*All*	*Vowel phonemes*	*Voiced segments*
30	5.22	3.61	3.22	37.22	34.98	30.12
36	4.12	2.82	2.78	36.12	32.95	29.11
42	3.28	2.37	2.12	35.21	31.54	29.01
48	3.31	2.04	1.98	33.84	29.84	28.43
54	2.82	1.62	1.52	33.01	29	28.02
60	2.12	1.36	1.12	32.89	27.34	26.81
66	2.28	1.08	1.10	31.44	26.35	24.12
72	1.90	0.96	0.92	30.08	25.52	25.31

Table 2. Speaker verification average error rates for variable amount of training data, for a population of 200 speakers, using all phonemes, vowel phonemes and voiced segments. Fixed test utterance length of 12 phonemes.

Number of Speakers	TIMIT Database IER(%)			NTIMIT Database IER(%)		
	All phonemes	*Vowel phonemes*	*Voiced segments*	*All phonemes*	*Vowel phonemes*	*Voiced segments*
50	1.88	0.81	0.78	30.22	23.21	21.02
100	2.12	0.81	0.79	31.28	23.41	21.53
150	2.52	0.93	0.85	31.90	23.47	22.12
200	2.33	1.01	0.91	32.12	23.51	22.59
250	2.98	1.04	0.93	32.58	24.24	23.10
300	3.01	1.07	0.96	33.55	24.73	23.41
350	3.22	1.03	0.98	33.98	25.03	23.78
400	3.34	1.05	1.00	34.12	25.55	23.91
450	4.12	1.07	1.02	36.22	26.27	24.11
510	6.22	1.14	1.13	38.88	26.75	25.08

Table 3. Closed-set identification error rate, as a function of the population size.

It can be observed that as the training data increase, the error rate decreases, reaching a minimum of 0.92% for TIMIT and 25.31% for NTIMIT at 72 voiced phonemes (216 speech frames) per speaker, which corresponds to a verification accuracy of 99.08% and 74.69% respectively.

For the identification experiments on the TIMIT and NTIMIT databases all 510 speakers (358 male and 152 female) were used. The goal of the experiments was to measure the performance of the system in closed-set and open-set identification for both clean and telephone speech. The closed-set identification accuracy as a function of population size for TIMIT and NTIMIT was tested. The corresponding identification error rates, given by the ratio of misclassifications (FC) to the total number of trials (TT), are shown in Table 3 in the column of voiced phonemes. It can be seen that increasing the population size ten times, the identification accuracy decreases by less than 0.35% for TIMIT and less than 4.06% for NTIMIT. We have also performed experiments with other phonetic categories, the results of which are shown in this table. We have trained again the speaker models by first using all the phonemes, then only vowels and finally only the voiced parts of the speech signal.

In open-set identification, where the classified speakers are further compared to decision thresholds, experiments were performed on both TIMIT and NTIMIT for 400 speakers (300 of them were members of the reference set and 100 were not). An acceptance error rate of 1.1% and a rejection error rate of 1.81% was obtained giving an average error rate of 1.46% for the TIMIT database. For the NTIMIT database an acceptance error rate of 27.12% and a rejection error rate of 27.92% was obtained giving an average error rate of 27.52%.

4. CONCLUSION

A text-independent speaker recognition system was presented based on neural networks. The proposed system detects and exploits the voiced segments of the speech signal which have the most speaker-dependent information. The system was tested for clean speech using the TIMIT database and for telephone speech using the NTIMIT database. It exhibited high accuracy rate, low response time, and low memory requirements (the reference data occupy less than 7.5 kbytes per speaker). The verification accuracy exceeded 99% for clean speech and 74.7% for telephone speech. The closed-set identification accuracy was measured to 99.22% for 50 speakers and 98.87% for 510 speakers for the TIMIT database, and 78.98% and 74.92% for the NTIMIT database respectively. The open-set identification accuracy was measured for 400 speakers to 98.63% for the TIMIT database and 72.39% for the NTIMIT. The proposed system is real-time and easy adaptable to new speakers since it requires very small training data.

The system was tested comparing the proposed approach of using voiced segments to other phonetic categories (only vowel phonemes and all phonetic categories). Voiced speech segments performed best in all tests.

5. REFERENCES

[1] Younès Bennani and Patrick Gallinari, "*Neural Networks for Discrimination and Modelization of Speakers* , Speech Communication, Vol. 17, pp. 159-175, 1995.

[2] P. Castellano and S. Sridharan, "*Speaker Identification with Projection Networks* , Proc. International Conference on Speech Science and Technology, Perth, pp. 400-405, December 1994.

[3] Jesper Olsen, "*Speaker Verification Based on Phonetic Decision Making* , in proc Eurospeech '97, Sept. 1997, pp. 1375-1378.

[4] J. Sirigos, N. Fakotakis, G. Kokkinakis: "*A comparison of several speech parameters for speaker independent speech recognition and speaker recognition* , in proc. Eurospeech '95, Madrid, Spain, 18-21 Sept. 1995.

[5] K.R. Farell and R.J. Mammone (1994), *Speaker recognition using neural networks and conventional classifiers* , IEEE Trans. Speech Audio Process., Vol. 2, No. 1, pp. 194-205.

[6] D. Anguita, M. Pampolini, G. Parodi , R. Zunino *YPROP: Yet Another Accelerating Technique for the Back Propagation* , in proc. ICANN '93, September 13-16 1993, Amsterdam, The Netherlands, p. 500.