FEATURE SELECTION FOR A DTW-BASED SPEAKER VERIFICATION SYSTEM

Medha Pandit and Josef Kittler

Centre for Vision Speech and Signal Processing
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey, Guildford GU2 5XH, UK
{j.kittler, m.pandit}@ee.surrey.ac.uk

ABSTRACT

Speaker verification systems, in general, require 20 to 30 features as input for satisfactory verification. We show that this feature set can be optimised by appropriately choosing proper feature subset from the input feature set. This paper proposes a technique for optimisation of the feature sets, in an Dynamic Time Warping (DTW) based text-dependent speaker verification system, to improve false acceptance rate. The optimisation technique is based on the l-r algorithm. The proposed scheme is applied to study cepstrum coefficients and their first order orthogonal polynomial coefficients. Experiments are conducted on two data bases: French and Spanish. The results indicate that with the optimised feature set the performance of the system may improve but it is never degraded. Moreover, the speed of verification is significantly increased.

1. INTRODUCTION

Speaker verification is the process of accepting or rejecting the identity claim of a speaker using speaker-specific information contained in speech signal. From this signal a set of features is extracted. Much research had been done on extraction of features from speech signal [1][14], which are useful for discrimination among speakers [16]. This feature set contains linguistic and speaker-dependent information.

Speaker related variations in speech are caused in part by the anatomical differences in the vocal tract and in part by the differences in speaking habits of different individuals. These variations are called inter-speaker variations but we must also consider intra-speaker variations-those occurring within different speech utterances of a single speaker [1]. The later variations are caused by many factors such as the differences in the speaking rates, the emotional state of speaker, his health etc. It is desirable to select for speaker recognition those acoustic parameters of speech which show low intra-speaker but high inter-speaker variability [13]. This issue is briefly discussed in Section 2. As we are interested in text-dependent verification, we adopt the Dynamic Time Warping matching algorithm described in Section 3, which in this context has been shown to outperform the Hidden Markov Model [8].

This paper addresses the problem of selecting discriminatory features from the input set of acoustic signal descriptors. This problem in context of speech recognition and speaker recognition has already been addressed in earlier studies. Cheung [5] proposed feature selection via dynamic programming for text-independent speaker identification. The paper also compares the “knock out” strategy with dynamic programming and shows that the identification error rate can be improved with proper selection of the feature set. Recently, Charlet [4] advocates the use of a different criterion function in conjunction with dynamic programming. The Speaker Verification used is based on the Hidden Markov Model approach. Torre and Peinado [15] proposed a new algorithm for feature selection based on the Discriminative Feature Extraction (DFE) technique and applied to speech recognition. The speech recognition feature extraction methods used now a days are not optimal when they are applied to specific environment and specific recognition task. With this motivation, Gu and Liu [9] proposed an optimisation extension to a previous non-optimal, universal feature extraction method.

Our work is distinguished from the earlier studies in two important respects. In the speaker verification scenario the speaker identity is known and therefore the acoustic features used for verification may be client dependent. Thus in our approach, the feature selection is user dependent. Furthermore, in contrast to Charlet [4], our feature selection process takes into account the effect of feature selection on warping. This in practice means that the time alignment function is optimised for each candidate feature set to evaluate its discriminatory effectiveness. In this sense our algorithm emulates the estimation-maximisation (EM) process where the steps of model selection and parameter estimation are alternated to find the optimal solution to the feature selection problem. The optimisation method of selecting a feature subset from input features is proposed in Section 4. It describes the l-r search algorithm [6], which minimises the experimental error rate in DTW-based speaker verification system. The proposed scheme is applied to study cepstrum coefficients and their first order orthogonal polynomial coefficients [7]. Experiments are conducted on two data bases and results are presented in Section 5. The results indicate that with the optimised feature set the performance of the system may improve but it is never degraded. Moreover, the speed of verification is significantly increased.

2. PARAMETER EVALUATION

Speaker identity is correlated with the physiological and behavioral characteristics of the speech production system for each speaker. These characteristics exist both in the spectral envelope (vocal tract characteristic) and in the supra-segmental features (voice source characteristic) of speech. It is impossible to separate these kinds of characteristics and are difficult to meas-

This work was partially supported by ACTS Project M2VTS

ure explicitly, hence many characteristics are captured implicitly by various signal measurements. Signal measurements such as short-term and long term spectra and overall energy are easy to obtain. These measurements provide the means for effectively discriminating among speakers [2] [10]. From these parameters, the selection of suitable speech attributes requires an appropriate criterion of effectiveness. For a single measurement parameter, this amounts to saying that a good measure of effectiveness would be the ratio of inter-speaker to intra-speaker variance, often referred to as the F ratio. A detailed discussion is given in [16][12] .

2.1. Feature Extraction

The measurements extracted from speech signal are cepstrum coefficients and their first order derivatives. Cepstrum coefficients are derived from the linear predictor coefficients. First tenth order linear predictor coefficients are extracted from each frame by the auto-correlation method. Then the linear predictor coefficients are transformed into cepstrum coefficient and orthogonal polynomial coefficients of cepstrum are calculated [7]. Here, we have used tenth order cepstrum coefficients and first order coefficient of time functions, which represents the slope of cepstrum. Thus a set of 20 features is used as input feature set.

3. VERIFICATION TECHNIQUE

The verification technique used is based on DTW. In this, time registration of the time functions of the sample utterance is made with the time functions retrieved as the reference template of the claimed identity. An overall distance between the sample utterance and the reference template is obtained as the result of time registration using the dynamic programming technique. The distance of each element is weighted by intra-speaker variability summed to produce the overall distance. Finally the best match distance is compared with a threshold distance value to determine whether the identity claim should be accepted or rejected [7]. The expression for the distance metric [7] adopted is:

$$D(R(n), T(m)) = \sum_{i=1}^{K} g_i^2 (r_i(n) - t_i(m))^2 \qquad (1)$$

where g_i is the weighting function, which is the reciprocal of the mean value of intra-speaker variability for the i^{th} element. Using this distance, the dynamic path is chosen to minimise the accumulated distance along the path.

3.1. Reference Pattern Construction

The procedure for establishing the initial reference template is the following. The first training utterance is used as a basic utterance, to which the second is brought into time registration. After registration the time functions of the feature parameters of first two utterances are averaged and the third is brought into time registration with the averaged function and then averaged into it. In the present case, four utterances are used as a basis for computing the reference template. So, the fourth is also brought into time registration and included in the averaging. The training utterances are also used for the calculation of the weighting function which is used in the distance measurement (see eq. 1).

3.2. Decision Threshold

The overall distance accumulated over the optimum warping function is compared with a threshold to determine whether to accept or reject an identity claim. To find a suitable threshold we measure the distances between the training utterances and the adopted template. The one which is largest is taken as the threshold.

The following section discusses the optimisation problem, involved in selecting an optimum feature set from the input feature set.

4. THE PROPOSED OPTIMISATION METHOD

We are interested in finding a subset of features which minimise the error rate of a speaker verification system. In this system, error rate depends on the decision threshold, hence we consider an empirical error rate (false acceptance rate) rather than its theoretical counterpart. To find an optimal set, is a combinatorial optimisation problem. The optimisation method can be specified in terms of two components:

(i) a performance criterion for the selection of optimum features from the input feature set.

(ii) optimisation procedure.

4.1. Feature Selection

4.1.1. General

The goal of feature selection is twofold: to reduce the dimensionality of the feature vector as required by any feasibility limitation of either technical or economical nature; to remove any redundant and irrelevant information, which may have a detrimental effect on the classifier performance. The problem of feature selection can be described as selecting the best subset X of d features, from the set Y,

$$X = \{x_i | i = 1, 2, 3....d, x_i \in Y\} \qquad (2)$$

$$Y = \{y_j | j = 1, 2, 3...D\} \qquad (3)$$

of $D > d$ possible measurements representing the pattern.

By best subset, we mean the combination of d features which optimises the criterion function $J()$, ideally the probability of correct classification, with respect to any other combination $\Xi = (\xi_i | i = 1, 2, 3...d)$ of d features taken from Y.

For the feature selection process, all the possible subsets of d out of D attributes should be considered to guarantee optimality of the feature set selected. The number of these sets is given by the well known combinatorial formula [6].

It is apparent that, even for moderate values of D and d, a direct exhaustive search will not be possible. Evidently, in practical situations, alternative, computationally feasible procedures will have to be employed. Such search algorithms, both optimal and suboptimal, that obviate the exhaustive search are discussed in [6]. The l-r algorithm is one of the suboptimal search algorithms mentioned in [6].

4.1.2. Search Algorithms for Feature Selection

(i) Sequential Forward Search (SFS)

It is the simple bottom up search procedure where one measurement at a time is added to the current feature set. The criterion function used for selection of feature is False Acceptance Error rate. At each stage, the attribute to be included in the feature set is selected from among the remaining available measurements (using the performance criterion), so that a new enlarged set of feature yields a minimum value of the criterion function used.

The algorithm is initialised by setting $X_0 = \phi$, where ϕ means null set [6].

(ii) Sequential Backward Search (SBS)

The SBS is the top down counterpart of the SFS method. Starting from the complete set of measurements, Y, we discard one feature at a time until $(D - d)$ measurements have been deleted. At each stage of the algorithm the element to be removed from the current feature set is determined by investigating the statistical dependence of the features in the set.

(iii) The l-r algorithm

Consider that we have input feature set Y and suppose k features have been selected to generate set X_k. l indicates the number of features to be added using SFS and r indicates the number of features to be discarded by the SBS method. In our work, we have used $l = 2$ and $r = 1$. The algorithm is described in steps as follows:

1. Using the SFS method add l features, ξ_j , from the set of available measurements, $Y - X_k$ to X_k, to create feature set X_{k+1}. Set $k = k + l$, $X_{D-k} = X_k$.
2. Remove the r worst features, ξ_j from the set X_{D-k} using the SBS procedure to form feature set X_{D-k+r}. Set $k = k - r$. If $k = d$ then terminate the algorithm. Otherwise set $X_k = X_{D-k}$ and return to step 1.

If $l > r$ then (l, r) algorithm is a bottom up search method. Commence from step 1 with k and X_0 set respectively to $k = 0$ and $X_0 = 0$. For $l < r$, the (l-r) algorithm is a top down procedure. Set $k = D$ and $X_0 = Y$ and start from step 2.

In all our experiments the above algorithms are used for optimisation of the input feature set.

5. EXPERIMENTS AND RESULTS

Experiments are conducted on two different data sets. One consists of 33 French male and female speakers of M2VTS data base [11] and other consists of 40 Spanish speakers [3]. In this DTW-based verification system, the utterance used for the experiment is a sentence of 0-9 digits spoken in French and Spanish. The model is trained using four repetitions of the same sentence spoken approximately at 1 week intervals. The features (cepstrum derived from LPC and orthogonal cepstrum) are averaged over the four repetitions and g_i (weighting function), which is a measure of intra-speaker variability, is also calculated recursively. Thus each utterance is transformed to speech features and weight (g_i) of each feature. Then the verification is performed using the Dynamic Time Warping (DTW) approach. For the feature selection, the l-r algorithm is used, which is described earlier. The performance criterion used for selecting features is %False Acceptance rate, as the False Rejection rate is 0% according to an adopted decision threshold strategy. Experiments are conducted separately on the French and Spanish databases.

For experimental evaluation, we have used the speech database, consisting of speech wave files obtained by sampling the waveform at 16 kHz and quantising each sample into 16-bit linearly. A high frequency emphasis filter is then applied to this digitised speech and a 30 ms Hamming window is used with 10ms overlap to extract the features (cepstrum derived from LPC and orthogonal cepstrum coefficients). There are 5 shots of the utterance for each speaker.

First experiment is conducted on the data base of 33 speakers (French). Each speaker is considered as a customer and others as impostors. The utterance used is segmented into two parts as: 0-3 (dataset-1) and 4-9 (dataset-2) for each customer and for each shot.

The experiments are done on these two data-sets:

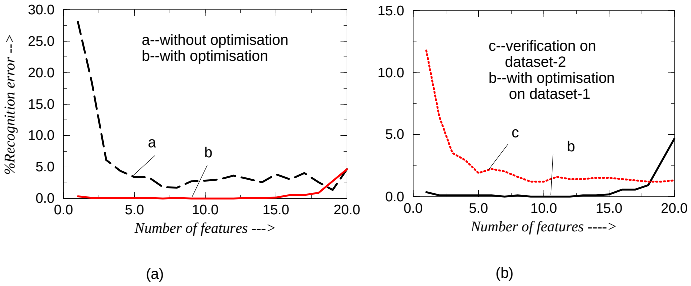

Figure 1: False Acceptance error rate obtained on data set-1.

A. In this experiment, the dataset-1 is used to train the model and the dataset-2 is used to evaluate weighting function for each feature. Then feature selection (l-r algorithm) is applied on dataset-1 using the trained model and an optimum feature set is obtained for each customer. The output results are shown in Fig.1. Fig.1(a) shows that by appropriately choosing proper feature set, the experimental error rate can be reduced down to about 0-1% depending on the number of features. It also shows that performance of the system deteriorates after the optimum set of 15 features and the error rate is 4.6% for 20 features. Taking this optimum feature set, the verification performance is tested on dataset-2 and its results are shown in Fig.1(b). For the optimum feature set of 10, the verification performance is 83.3%, which is the same when all 20 features were used on dataset-2. This shows that verification can be carried out using a subset of acoustic feature without degrading the performance of the system. The recognition error (%FA) at this optimum feature set is 1.2%.

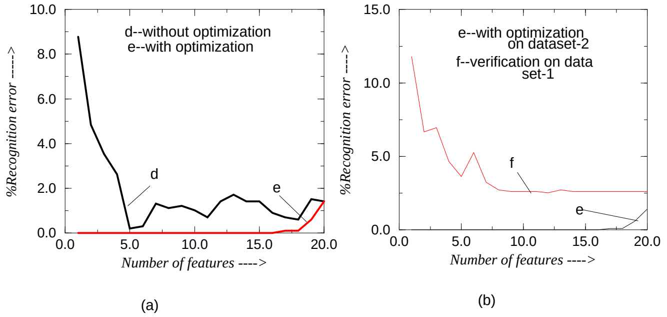

Figure 2: False Acceptance error rate obtained on data set-2.

B. In this experiment, the dataset-2 is used to train the model and dataset-1 is used to evaluate the weighting function for each feature. Then the l-r feature selection algorithm is applied on dataset-2 using the trained model and an optimum feature set is obtained for each customer. The output results are shown in Fig.2. Fig.2(a) shows that the performance deteriorates after optimum set of 16 features and the error rate is 1.3% for 20 features. This optimum feature set is then used for verification on dataset-1 and the results are presented in Fig.2(b). For optimum feature set of 13, the verification performance is 78.8% , which is same when all 20 features are used for verification on dataset-1. The recognition error (%FA) at this optimum feature set is 2.5% and this shows an improvement in this rate , as it is 4.6% with 20 features (see Fig.1(a)).

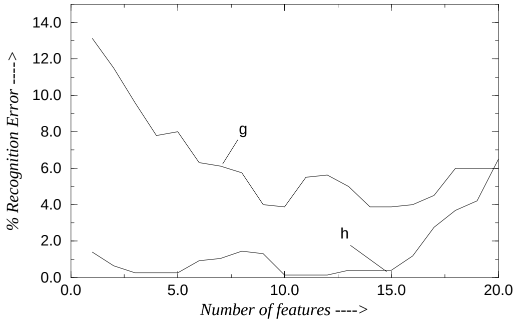

Figure 3: False Acceptance error rate obtained on Spanish data.

The other experiment is conducted on Spanish data base of 40 customers. In this experiment, 40 speakers are used as customer and 20 as impostors. For each customer the imposter set is different. There are 6 shots for each customer. Shots 1-4 are used to train the model and a different utterance containing the name and address of each customer is used to evaluate the weighting functions for each feature. Then feature selection is applied on this trained model along with shot 5 and the optimum feature set is obtained. The results are shown in Fig.3. Graph h shows the outcome of the feature selection process and graph g shows the verification results using shot 6 for testing with the optimum feature sets on the model trained earlier. The %FA rate at optimum feature set of 10 is 3.87% as compared to 6% for all 20 features ,which shows a significant improvement in error rate .

These experiments show that by optimising the set of acoustic features using the feature selection technique, the verification error rate can be significantly reduced in addition to increasing the speed of processing. The optimum feature set which we get from this experiment, mostly contains first order cepstrum coefficients and higher order cepstrums for all customer. This shows that for text-dependent speaker verification, much of speaker-dependent information is contained in transitional coefficients.

6. CONCLUSION

In this paper, we have addressed the problem of optimising the acoustic feature set for text-dependent speaker verification, in a Dynamic Time Warping system. We applied the l-r feature selection algorithm to study cepstrum coefficients and their first order derivatives. The experimental results on French database show that an optimum feature set can be obtained without degrading the performance of the system, while the experiment on Spanish data shows a significant improvement of verification error rate. The experiments also highlighted the usefulness of first order orthogonal cepstrum coefficients and higher order cepstrum.

Further improvements in performance of the system are expected from combining speech and lip features and from optimising the bimodal feature sets.

7. REFERENCES

[1] B. Atal. Automatic recognition of speakers from their voices. In *Proceeding of IEEE, vol.64, No.4*, pages 460–475, 1976.

[2] M. J. Carey and E. S. Parris. Robust prosodic features for speaker identification. In *Proc. Int.Conf. Spoken Language Processing, Philadelphia*, pages 1800–1803, 1996.

[3] CARLOS. Speech database, Universidad Carlos III de Madrid, Spain, 1996.

[4] D. Charlet and D. Jouvet. Optimizing feature set for speaker verification. In *Proceedings of First International Conference, AVBPA'97*, pages 203–210, 1997.

[5] R. Cheung and B. Eisenstein. Feature selection via dynamic programming for text-independent speaker identification. In *IEEE Trans. on Acoust.,Speech and Signal Processing, vol. ASSP-26, NO.5*, pages 397–403, 1978.

[6] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, Englewood Cliffs, NJ, 1982.

[7] S. Furui. Ceptral analysis technique for automatic speaker verification. In *IEEE Transactions on Acoustics, Speech and Signal Processing, vol.2*, pages 254–272, 1981.

[8] S. Furui. Recent advances in speaker recognition. In *AVBPA97*, pages 237–251, 1997.

[9] L. Gu and R. Liu. The application of optimization in feature-extraction of speech recognition. In *International Conference on Signal Processing Proceedings,ICSP, vol.1*, pages 745–748, 1996.

[10] T. Matsui and S. Furui. Text-independent speaker recognition using vocal tract and pitch information. In *Proc. Int.Conf. Spoken Language Processing, Kobe, 5.3*, pages 137–140, 1990.

[11] S. Pigeon and L. Vandendrope. The M2VTS multimodal face database (release 1.00). In *AVBPA97*, pages 403–409, 1997.

[12] S. Pruzansky and M. Mathews. Talker recognition procedure based on analysis of variance. In *Journal Acoustic Society America, vol.36*, pages 2041–2047, 1964.

[13] A. Rosenberg. Automatic speaker verification: A review. In *Proceeding of IEEE, vol.64, No.4*, pages 475–487, 1976.

[14] M. Sambur. Selection of acoustic features for speaker identification. In *IEEE Transactions on Acoustic, Speech, and Signal Processing, vol.ASSP-23*, pages 176–182, 1975.

[15] A. Torre and Antonio M. Peinado. A def-based algorithm for feature selection in speech recognition. In *Proc. Int.Conf.on Acoustic,Speech and Signal Processing,Germany,vol.II*, pages 1519–1522, 1997.

[16] J. Wolf. Efficient acoustic parameters for speaker recognition. In *Journal Acoustic Society America,vol.51*, pages 2044–2055, 1972.