Selection of Speaker Independent Feature for a Speaker Verification System

M. Pandit, J. Kittler and J. Matas
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, g.matas}@ee.surrey.ac.uk

Abstract

In this paper we propose an optimisation technique to choose a user independent feature subset from the input feature set for a DTW-based text-dependent speaker verification system. The results indicate that with the optimised feature set the verification error rate of the system can be improved.

1. Introduction

Speaker verification is the process of accepting or rejecting an identity claim of a speaker using speaker-specific information contained in speech signal. From this signal a set of acoustic descriptors is extracted. Much research had been done on extraction of features from speech signal [11][12], which are useful for discrimination among speakers [14] and should contain linguistic and speaker-dependent information.

As we are interested in *text-dependent* verification, we adopt the Dynamic Time Warping matching algorithm described in Section 2, which in this context has been shown to outperform the Hidden Markov Model [7].

This paper addresses the problem of selecting discriminative features from the input set of acoustic signal descriptors. This problem in the context of speech recognition and speaker recognition has already been addressed in earlier studies [4] [3] [13].

In our work, the feature selection process is user independent as opposed to the previously investigated user dependent approach [9]. In the user dependent case each user has its own feature subset for verification, while in the user independent one there is only one common feature subset for all the clients. Thus the user-independent approach has immediate merit over the client dependent counterpart when the number of client increases. Furthermore, in contrast to Charlet [3], 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 discriminative 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 3. It describes the l-r search algorithm [5], which minimises the experimental error rate in DTW-based speaker verification system. The proposed scheme is applied to cepstrum coefficients and their first order orthogonal polynomial coefficients [6]. Experiments are conducted on a Spanish database [2] and results are presented in Section 4.

2. Verification Technique

The measurements extracted from speech signal are cepstrum coefficients and their first order orthogonal polynomial coefficients. 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 coefficients and finally orthogonal polynomial coefficients of the cepstrum are calculated [6]. Here, we have used tenth order cepstrum coefficients and first order coefficients of their time functions, which represent the slope of the cepstrums. Thus a set of 20 features is used as an input feature set. These measurements have been shown to contain information for discriminating among speakers [1] [8].

The verification technique used is based on DTW. Accordingly, 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 a result of the time registration us-

[*] This work was partially supported by EU Project M2VTS.

ing 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 [6]. The expression for the distance metric [6] adopted is:

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

where g_i is the weighting function, which is the reciprocal of the mean value of intra-speaker variability for the i^{th} element. The $R(n) = (r_1(n)...r_K(n))$ and $T(m) = (t_1(m)...t_K(m))$ are the reference and test template feature vector of n^{th} and m^{th} frame of speakers respectively and K is the number of elements of feature vector. Using this distance, the dynamic path is chosen to minimise the accumulated distance along the path.

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.

3. The Proposed Optimisation Method

We are interested in finding a subset of features which minimise the error rate of our speaker verification system. This contrasts with previously reported work [4][12] where, a theoretical criterion function was used as a measure of effectiveness. 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.

Formally the problem of feature selection can be described as selecting the best subset X of d features, from the set Y,

$$X = \{x_i | i = 1, 2, 3....d, x_i \in Y\} \quad (2)$$
$$Y = \{y_j | j = 1, 2, 3...D\} \quad (3)$$

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

By best subset, we mean the combination of d features which optimises the criterion function 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 [5]. 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. The l-r algorithm is one of the suboptimal search algorithms mentioned in [5]. We are not using its more advanced versions [10] for computational reasons.

Search Algorithms for Feature Selection

Sequential Forward Search (SFS) 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 the null set [5].

Sequential Backward Search (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.

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 the (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.

4. Experiments and Results

Experiments are conducted on a Spanish data set of 40 speakers [2]. In this DTW-based verification system, the utterance used for the experiment is a sentence of 0-9 digits spoken in Spanish. The model is trained using four repetitions of the same sentence spoken approximately at 1 week

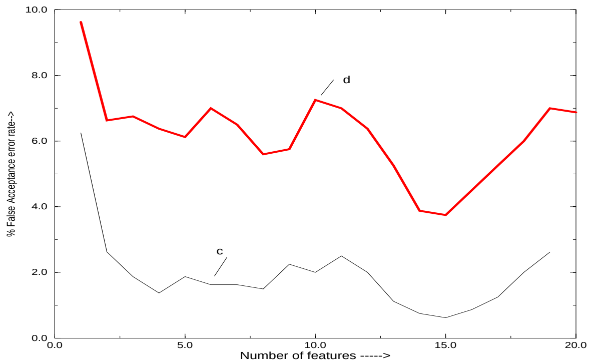

Figure 1. False Acceptance error rate.

intervals. The acoustic descriptors (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(FA) rate, as the False Rejection(FR) rate is 0 according to an adopted decision threshold strategy.

In the experiment, the following procedure is repeated for each client of the database:- Let speaker x is used as client. From the remaining 39 speakers of data base, 20 speakers are used as impostors excluding client. For the client x, shots 1-4 are used to train the model and shot 5 of 20 impostors is used in feature selection process and feature subsets are obtained. Then verification is performed using shot 6 and the obtained feature subsets. In verification one client test and 19 impostors tests are performed. The set of 19 impostors is different than the one used in the feature selection process. A different utterance containing the name and address of client x is used to evaluate the weighting functions for each feature.

The results are shown in Fig. 1. Graph c shows the outcome of the feature selection process and graph d shows the verification results using shot 6 for testing with the optimum feature sets of different cardinality on the model trained earlier. The FA rate at optimum feature set of size 15 is 3.7% as compared to 6.9% 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. From [9], the number of speaker independent features required to achieve a comparable performance to the speaker dependent approach is 50% higher. However it may be beneficial to accept this increase for the sake of simplicity of the verification system.

5. Conclusion

In this paper, we have addressed the problem of optimising the acoustic feature set for text-dependent speaker verification, using a Dynamic Time Warping system. We applied the l-r feature selection algorithm to study the effectiveness of cepstrum coefficients and their first order derivatives and to select user independent feature subset. The experiment on Spanish data shows a significant improvement of verification error rate with optimum feature set.

References

[1] 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.

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

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

[4] 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.

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

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

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

[8] 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.

[9] M. Pandit and J. Kittler. Feature selection for a DTW-based speaker verification system,. In *Proceeding of ICASSP'98, vol. 2*, pages 769–772, 1998.

[10] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. In *Pattern Recognition Letters, 15*, pages 1119–1125, 1994.

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

[12] 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.

[13] A. Torre and A. 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.

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