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 optimisation technique is based on the l-r algorithm, which in essence is the combination of sequential forward and backward search algorithms. The performance criterion used for optimum feature selection is the experimental error rate. The proposed scheme is applied to cepstrum coefficients and their first order orthogonal polynomial coefficients. Experiments are conducted on a data base of 40 Spanish male and female speakers. The results indicate that with the optimised feature set the verification error rate of the system can be improved. Moreover, the speed of verification is significantly increased.

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 [12][13], which are useful for discrimination among speakers [15] and should contain linguistic and speaker-dependent information.

Speaker related variations in speech are caused in part by anatomical differences in the vocal tract and in part by 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 [12]. The later variations are caused by many factors such as the differences in the speaking rates, the emotional state of speaker, his health etc. For speaker recognition, it is desirable to select those acoustic parameters of speech which show low intra-speaker but high inter-speaker variability [12]. 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 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. Cheung [4] 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 [3] advocated the use of a different criterion function in conjunction with dynamic programming. The speaker verification method used is based on the Hidden Markov Model approach. Torre and Peinado [14] proposed a new algorithm for feature selection based on the Discriminative Feature Extraction (DFE) technique and applied to speech recognition.

In our work, the feature selection process is user independent as opposed to the previously investigated user dependent approach [10]. 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 4. 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 [7]. Experiments are conducted on a Spanish data base [2] and results are presented in Section 5. The results indicate that with the optimised feature set the performance of the system can be improved. 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 of each speaker. These characteristics influence both the spectral envelope (vocal tract characteristic) and the suprasegmental features (voice source characteristic) of speech. It is impossible to separate these kinds of characteristics which are difficult to measure 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 have been shown to contain information for discriminating among speakers [1][9]. However, in order to identify which of these parameters are most effective, an appropriate criterion of effectiveness is required. In previously reported work [4][13], a theoretical criterion function was used as a measure of effectiveness. In this paper we adopt an empirical error rate as a criterion for selecting the most discriminatory acoustic features.

2.1. Acoustic Descriptor Extraction

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 [7]. 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.

3. Verification technique

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 result of the time registration using 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:

\[ D(R(n), T(m)) = \sum_{i=1}^{K} g_i^2(r_i(n) - t_i(m))^2 \]  (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 \( R(n) = (r_1(n) \ldots r_K(n) \) and \( T(m) = (t_1(m) \ldots t_K(m)) \) 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.

3.1. Reference Pattern Construction

The procedure for establishing the initial reference template for each client 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 in (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 our 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. Selection of the 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

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 \( Y \),

\[
X = \{x_i|i = 1,2,3,...d, x_i \in Y\} \tag{2}
\]

\[
Y = \{y_j|j = 1,2,3,...D\} \tag{3}
\]

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 \([5]\). 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. Such search algorithms, both optimal and suboptimal, that obviate the exhaustive search are discussed in \([5]\). The \( l-r \) algorithm is one of the suboptimal search algorithms mentioned in \([5]\). We are not using its more advanced versions \([11]\) 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 \( \phi \) 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, \( \xi_j \), 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, \( \xi_j \), 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.

5. 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 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 \( g_i \) (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.

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 emphasised 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) [6]. There are 6 shots of the utterance for each speaker.

In the experiment of feature selection and verification, following procedure is repeated for each client of data base:

Let speaker $x$ be 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 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(b). Graph $e$ 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. Fig.1(a) shows the results of the same experiment with user dependent feature sets[10]. The FA rate in this case is 3.87% for an optimum feature set of size 10 as compared to 6% for all 20 features.

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 Fig.1(b) contains:

$$\delta f = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, f_1, f_3, f_5, f_6, f_7, f_9\}$$

where $c_1, c_2...$ are cepstrum coefficients and $f_1, f_3..$ are first order orthogonal polynomial coefficients of cepstrum. From Fig.1, 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 and some savings in storage capacity that would be needed to store the indices of user dependent attributes.

6. 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 1-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


