Speaker Verification by Means of ANNs

Urs Niesen and Beat Pfister

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland

{niesen,pfister}@tik.ee.ethz.ch

Abstract.
In text-dependent speaker verification the speech signals have to be time-aligned. For that purpose dynamic time warping (DTW) can be used which performs the alignment by minimizing the Euclidean cepstral distance between the test and the reference utterance. While the cumulative Euclidean cepstral distance, which can be gathered from the DTW algorithm, could be used directly to discriminate between a pair of signals spoken by the same and by two different speakers, we show that a distance measure learned by an artificial neural network performs significantly better for the same task.

1 Introduction

The problem of speaker verification consists in either accepting or rejecting a claimed identity based on a test signal recorded from the person to be authenticated and a reference signal corresponding to the claimed person. Inter-speaker variations, which may be exploited for this task, can be caused by physical differences of the vocal tract or by different speaking habits such as rhythm and intonation or dialects. Intra-speaker variations on the other hand are induced for example by the emotional state of the speaker, his health or aging [1].

Two general approaches to speaker verification are possible. Either the text to be uttered is prescribed or not (referred to as text-dependent vs. text-independent speaker verification). The motivations for a text-dependent, pattern matching-based approach [1] are mainly the good performance for short utterances and the language-independency. Effectively the language does not matter. Reference and test signals are just required to have identical wording.

Even though the uttered text is the same for text-dependent systems, the speech signals will in general not be aligned due to variability inherent to natural speech. Some kind of time alignment has therefore to be performed prior to comparison. This can be done by means of dynamic time warping (DTW). Once this alignment

---

[1]There exist also text-dependent systems which are not based on pattern matching, such as [3].
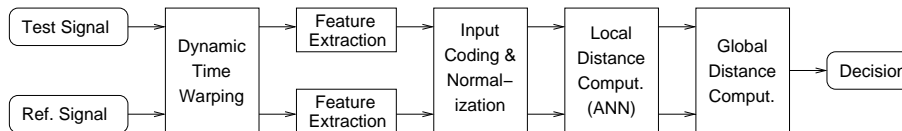
Figure 1: SV system with ANN-based local distance computation

is achieved, text-dependent systems have the big advantage that corresponding frames of the reference and the test signal can directly be compared to each other, i.e., for each pair consisting of a reference and a test frame a distance (the so-called local distance) is computed.

The global distance, based on which the final decision will be made, can be obtained by averaging the local distances over the length of the signal. This approach is similar to the one taken for example in [4], which uses the Euclidean cepstral distance[2] in the DTW and also to compute the global distance for the discrimination. The requirements for a distance metric for these two tasks are however very different. For the DTW a metric is needed which separates different from similar sounds. For the discrimination task a metric is needed which separates speech spoken by the same from speech spoken by different speakers. It seems unlikely that the same distance metric is optimal for both tasks.

In this study the use of an alternative ANN-based distance measure for the discrimination task is investigated. More precisely, the ANN replaces the Euclidean distance computation only and thus uses as input pairs of feature vectors that have been extracted from pairs of aligned speech frames.[3]

Preliminary investigations of various feature types indicated, that the LPC cepstrum works best with the proposed method for speaker verification, which coincides with the findings for example in [2].

In the next Section the application of the ANN is described. In Section 3 the results are discussed and compared to a system using Euclidean cepstral distance for the discrimination task. In Section 4 some concluding remarks are given.

2 System description

The block diagram in Figure 1 shows the main operations of the ANN-based speaker verification system. The two speech signals are time-aligned with the DTW algorithm that uses Euclidean cepstral distance as optimization criterion (the preprocessing steps, i.e., band-limiting, windowing and cepstral mean subtraction are not shown). For each frame of the aligned signals, the features are extracted, which in this case are the 12 first coefficients of the LPC cepstrum. The feature vectors of two corresponding frames, one from the reference and one from the test signal, are then combined (see section 2.2) and fed in the ANN which yields a local distance for this pair of

---

[2]The mean square difference between two smoothed logarithmic spectra is efficiently expressed as Euclidean distance of the respective time-limited cepstra.

[3]Basically the ANN could replace both the feature extraction and the distance computation. In this case the ANN would directly process pairs of speech frames. This approach has not yet been considered.

frames. The local distances are weighted[4] and averaged over time. The result is called the global distance for the whole speech signal, which is compared to a predetermined threshold for classification.

2.1 ANN configuration

The ANN used in this study is a fully connected multi-layer perceptron with hyperbolic tangent activation function. For the training the well-known back-propagation algorithm was applied together with adaptive learning rate. The weights were updated after each epoch as follows:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \boldsymbol{\Delta w}(t) \tag{1}$$

$$\boldsymbol{\Delta w}(t) = -\mu(t)\frac{\partial E}{\partial \boldsymbol{w}}(t) + \alpha(t)\,\boldsymbol{\Delta w}(t-1) \tag{2}$$

$$\mu(t) = \begin{cases} 1.05\,\mu(t-1) & \text{if } E(t) < E(t-1) \\ 0.7\,\mu(t-1) & \text{if } E(t) > 1.01\,E(t-1) \\ \mu(t-1) & \text{otherwise} \end{cases} \tag{3}$$

$$\alpha(t) = \begin{cases} 0.99 & \text{if } E(t) < 1.01 E(t-1) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $E(t)$ is the MSE between the desired output (0 for identical and 1 for different speakers) and the effective output of the ANN at epoch t.

As mentioned above, the ANN is used to compute a distance between a pair of feature vectors. A feature vector consists of the cepstral coefficients $c(1)\ldots c(12)$, and thus 24 input nodes and 1 output node are needed. The number of hidden layers and nodes were determined experimentally. Optimal results were found with 2 hidden layers having 60 nodes in the first and 18 nodes in the second hidden layer.

2.2 Input coding and normalization

Ideally the output y of the ANN should behave like a metric and particularly should have the following symmetry property: $y(\boldsymbol{x}_1, \boldsymbol{x}_2) = y(\boldsymbol{x}_2, \boldsymbol{x}_1)$, i.e., the distance should be the same if the feature vectors of the reference and the test frame are exchanged. In principle the ANN could learn this property from appropriately designed training data. It has turned out, however, that the training of the ANN is extremely difficult and slow with such input data.

In order to build the desired invariance into the system, we defined a function $g(\boldsymbol{x}_1, \boldsymbol{x}_2) = g(\boldsymbol{x}_2, \boldsymbol{x}_1)$ which has the symmetry property mentioned above, and use the coded input $g(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to train the ANN (see also Figure 1). Several reasonable functions $g(.,.)$ exist. The one which yielded the best result is

$$g(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 + \boldsymbol{x}_2, |\boldsymbol{x}_1 - \boldsymbol{x}_2|) \tag{5}$$

This function computes the sum and the absolute difference between the feature vectors, and therefore has the desired symmetry property. Since the absolute difference

---

[4]In noisy signals, frames with low energy are dominated by the noise characteristics rather than by speech. A weighting which accounts for this is the sum of the square roots of the test and the reference frame energies.

can be expected to be an important information for the classifier, it potentially facilitates also the training of the ANN. Indeed this type of input coding reduced the number of training epochs dramatically.

In addition to this coding, the input of the ANN was normalized by means of a linear transformation in order to have zero mean and diagonal covariance matrix. This transformation has been found with principal component analysis (PCA). As can be seen from Figure 2 this normalization again improved the training.
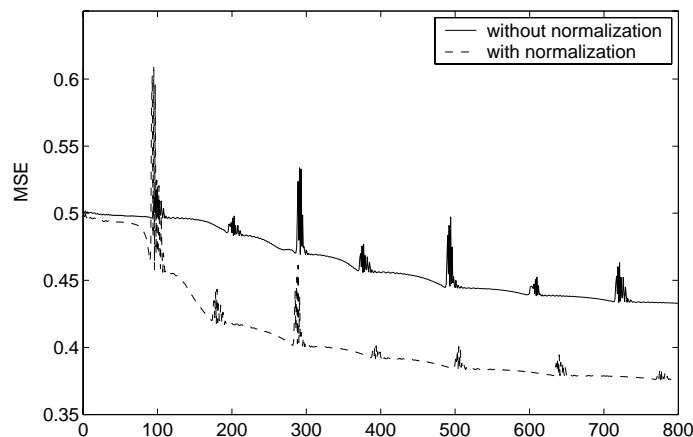

Figure 2: Learning curves (MSE of local distances in function of training epochs) for the validation error for the same ANN with and without normalization

2.3   Using transitional information

In a first approach to using transitional information, simply the first and the second temporal derivatives of the cepstral coefficients were additionally fed to the ANN. For that purpose the number of input nodes had to be tripled. Indeed the Fisher ratio [5] for the local distances increased considerably. The global distances however showed worse performance. It seemed surprising at first, that the use of additional information can result in a decreased performance of the system. This fact can however be explained by recalling that the derivatives are estimated by means of first and second order regression over some 150 ms and 250 ms, respectively (see for example [5]). These estimates are called delta and delta-delta cepstrum.

The averaging operation performed during regression has two important consequences. First, the longer the regression window the better the local distances get (due to noise reduction). Second, the correlation of consecutive local distances based on longer regression windows is higher, which reduces the contribution of information to the global distances.

As explained above, the ANN is trained to optimally evaluate the local distances. On the frame level the derivatives have a higher discriminative capability than the

---

[5]From means and variances of two classes the Fisher ratio is evaluated as $\sqrt{(m_1 - m_2)^2/(\sigma_1^2 + \sigma_2^2)}$.

cepstral coefficients themselves, and the training will therefore put more emphasis on transitional than on instantaneous features. This will however yield suboptimal global distances.

In principle it would be possible to optimize the training for a block of consecutive frames that is large enough to neutralize or at least reduce the effect of the stronger correlation of the delta features. We preferred an alternative solution, however, namely to use three parallel ANNs for the classification, one for the instantaneous features and one for each type of transitional features. The local distance results from a linear combination of the ANN outputs. The optimal weights were found by means of discriminant analysis which has to be performed on the global distances.

For the output of the ANNs using cepstral, delta and delta-delta cepstral features, the optimal combination weights were found to be 0.86, 0.46, and 0.22, respectively. This shows that instantaneous features do bear more discrimination information than transitional features, but also that a combination of both is useful.

3   Experimental Results

The speech signals used in this study were collected from 30 male speakers. Some 10 sessions per speaker were recorded within a timeframe of several months. Each speaker used different telephones and the signals were directly recorded from the digital telephone network in a-law format. The total duration of the collected speech is about 3 hours. The features (LPC cepstral coefficients) were extracted from 37.5 ms long speech frames with a frame shift of 15 ms.

Each of the ANNs was trained with some 500,000 feature pairs from 20 speakers, about half for each class (i.e., identical or different speakers in pair). Tests were made with speakers from the training set and with the test speakers, i.e. the 10 remaining speakers. Thereby the ANNs showed good generalization and therefore can be considered as virtually speaker-independent.

Figure 3 shows the receiver operating characteristics (ROC) for three different methods of local distance evaluation in the discrimination task: Euclidean cepstral distance, ANN-based cepstral distance and the method with three parallel ANNs. As can be seen from this Figure the gain in performance is quite substantial, mainly due to the ANN. The equal error probability (EEP) drops from 9.1 % to 5.3 % for roughly one second long speech signals from the 10 test speakers.

4   Conclusions

In this work an alternative ANN-based metric to discriminate between utterances from the same and from different speakers was presented. The results were compared to a system, which uses the Euclidean cepstral distance as metric for the same purpose. The improvements by the proposed method are quite high. Further improvement was achieved by including transitional information in the form of the delta and delta-delta cepstral coefficients. In total the new approach reduced the equal error rate to almost half of the original system.
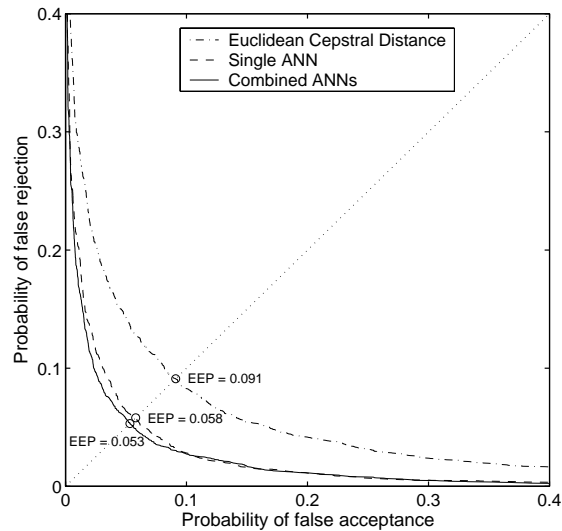
Figure 3: ROC curves for about 1 s long speech signals using three types of local distance evaluation

A further important result is the very good generalization of the ANN-based metric for unseen speakers. The ANN seems not to learn speaker-specific features to distinguish them, but rather a general rule. This is a very important aspect for practical applications: New speakers do not require a retraining of the ANNs.

Acknowledgement

References

[1] B. S. Atal. Recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4):460–475, April 1976.

[2] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, June 1974.

[3] M. BenZeghiba and H. Bourlard. *Hybrid HMM/ANN and GMM Combination for User-Customized Password Speaker Verification*. IDIAP research report 02-45, November 2002.

[4] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on ASSP*, 29(2):254–272, 1981.

[5] J. S. Mason, X. Zhang: *Velocity and Acceleration Features in Speaker Recognition*, Proceedings of ICASSP'91, pages 3673–3677, April 1991.