Speaker Verification Using Artificial Neural Networks

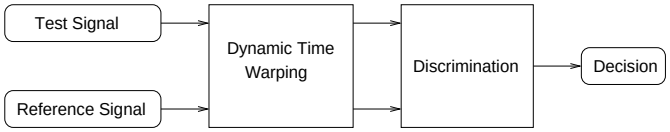
URS NIESEN AND BEAT PFISTER
Speech Processing Group, Computer Engineering and Networks Laboratory, ETH Zurich

Summary

A combination of artificial neural networks (ANN) can be used to discriminate between a pair of speech signals uttered by the same and by different speakers. Our experiments show that a significant improvement of the classifier can be achieved with this new approach compared to the Euclidean cepstral distance.

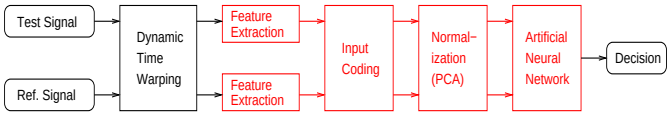
Motivation

For language-independent SV, we use a text-dependent system which relies on pattern matching. Two identically worded speech signals are time-normalized and the weighted average of the Euclidean cepstral distance (ECD) between the corresponding frames is used to discriminate between the two hypotheses "same speaker" and "different speakers".



- DTW is performed with ECD as distance measure.
- Speaker discrimination is based on ECD as well.

The distance measure for DTW has to differentiate between similar and different phones. The metric for speaker discrimination must distinguish between the same and different speakers. These quite different requirements make it unlikely, that the ECD is optimal for both.

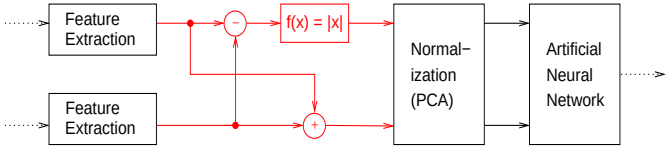


In order to improve the discrimination between speakers, a distance measure learned by an ANN has been used instead of ECD to evaluate the local distances between frames.

- What features are optimal (instantaneous and transitional)?
- Which invariances can be built into the system?
- How does the system generalize for speakers not in training set?

Input Coding

Ideally the function estimated by the ANN should behave like a metric. Most important is the symmetry property, which can be built into the system by coding the input data.



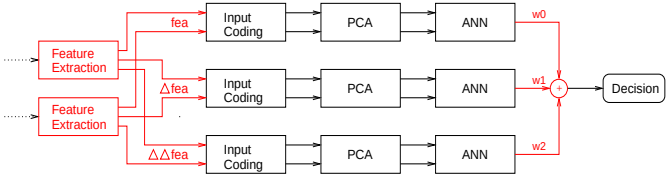
This guarantees the desired invariance. Moreover it speeds learning, as providing the absolute value of the feature difference to the ANN makes discrimination easier.

Using Transitional Information

Several features such as ACF, LPC, CEP etc. and their temporal derivatives have been investigated. Transitional features can be derived by regression over the instantaneous features of multiple frames. From this the derivatives are estimated. The averaging operation performed for the regression has two effects:

- Transitional features provide better discrimination for local distances (i.e. at frame level).
- The higher correlation between them makes them perform worse for global distances.

Therefore the instantaneous and transitional features cannot be given to the same ANN, as it would optimize the local and not the global distances. The global distance is relevant for the decision, however.

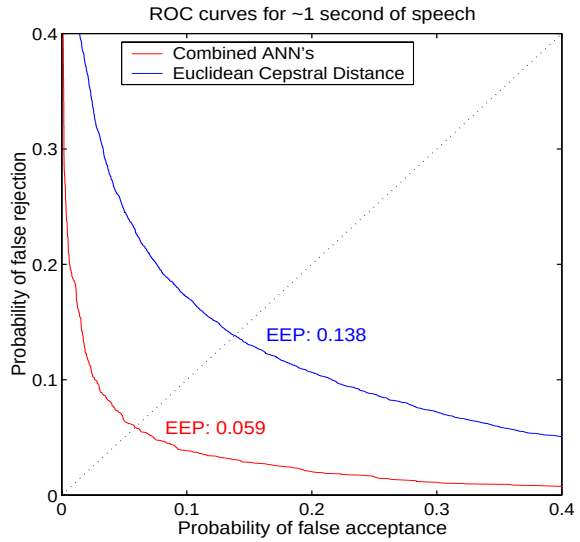


The weight vector $\boldsymbol{w}$ can be found e.g. using discriminant analysis. For cepstrum, delta and delta-delta cepstrum the optimal linear combination is found to be $\boldsymbol{w} = (0.86, 0.46, 0.22)^T$, showing that transitional features bear less discrimination information but also that a combination increases system performance.

Results

- The ANN was trained with about five hundred thousand frame comparisons from 20 different speakers. Good generalization capability was observed for speakers not present in the training set.
- Best results were achieved with cepstral features (cepstrum, delta and delta-delta cepstrum).

The receiver operating characteristic (ROC) diagram compares the performance of the original system (blue curve) to the new system using ANNs (red curve). The equal error probability (EEP) drops from 13.8% to 5.9% for roughly one second long speech segments recorded over telephone line. This corresponds to an improvement of Fisher ratio from 1.44 to 2.26.



Conclusion

Considerable performance gain is achieved by the proposed method. While the training time is pretty big (several days until convergence to a local minimum), the classification time is short. Also as the ANNs are not specific to each speaker, the system can be used for new speakers without having to retrain the classifier. This makes an application of the proposed system well possible for a variety of different situations.