2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94) Kyoto Research Park, Kyoto, Japan September 26 - 27, 1994

AN EXPERIMENTAL COMPARISON OF DIFFERENT FEATURE EXTRACTION AND CLASSIFICATION METHODS FOR TELEPHONE SPEECH

Tilo Schürer

Technische Universität Berlin
Institut für Fernmeldetechnik
Einsteinufer 25
D - 10587 Berlin
Germany
tilo@cs.TU-Berlin.DE

ABSTRACT

Robust speech recognition over telephone lines severely depends on the choice of the feature extraction and classification methods. In order to get the highest possible performance of the speech recognizer a number of commonly used feature extraction methods (MFCC, LPC, PLP, RASTA-PLP) *and* classification methods (MLP, LVQ, HMM) were tested on the same telephone speech data. All combinations of feature extraction and classification methods were computed and several parameters of both methods where changed in order to find a non-local maximum of recognition accuracy. The paper does not describe a comparison of classification but of feature extraction methods because it is clear that an HMM would outperform both LVQ and MLP. The big question is if the same feature extraction methods always lead to the best results, no matter which classifier is used!

Keywords: Feature extraction, classification, speech recognition.

1. INTRODUCTION

State of the art HMM recognizers for telephone speech use LPC- or MFCC-analysis [1], [2], [3]. In 1990, Hermansky first introduced a new feature extraction method, called *Perceptual Linear Predictive (PLP)* - analysis [4]. He showed that PLP clearly outperforms the LPC in speaker-independent mode even when using a lower model order. Later, Hermansky extended the PLP analysis by means of filtering and adaptation to the communication channel, and he showed that this approach is very robust against noise [5], [6]. This method is called *RelAtive SpecTrAl* (RASTA)-PLP.

When comparing different feature extraction methods with each other, very often only *one* classification method is used. This implies that those feature extraction methods would give the same performance when the classifier is replaced by another. This strategy could be observed, too, p.e. when evaluating the performance of *different* classifiers with *one* feature extraction method.

The question arises if it is really valid to reduce the high-dimensional parameter search-space in the described manners in order to find the best local maximum of recognition performance. In other words, shouldn't new feature extraction methods always be tested against *many* classifiers and shouldn't the same be done when testing new classification methods?

Trying to answer that question, an experimental comparison of different feature extraction methods (MFCC, LPC, PLP, RASTA-PLP) in combination with different classification methods (MLP, LVQ, HMM) was done. Since it is impossible to test the entire high-dimensional search-space, the performance of several combinations of both feature *and* classification methods is compared using a telephone speech database recorded in the area of Berlin, Germany. The aim of this paper is to show, whether the above mentioned reduction of search-space, which is usally done, really is valid and usable.

2. EXPERIMENT DESCRIPTION

2.1. Speech Database

The speech database was recorded over the public switched telephone network in the area of Berlin, Germany. It consists of the German isolated digits from zero to nine and 5 additional command words that are necessary to build simple voice-activated information- or voice-mail services. The data was recorded under noisy conditions (p.e. from public phones in streets, phones in buildings, offices with many people and computers running) in order to get *realistic* data. The speech was recorded via an ISDN-card on an Intel 486 PC running the UNIX-System *LINUX*. Most of the calls were transmitted over analogue telephone lines, only approximatly 10% were recorded over ISDN or digitally switched lines. A total of 250 speakers was recorded[1]. All speech data was automatically endpointed using an energy-based method after Savoji [7]. Randomly selected 80% of all digits formed the training set, while the remaining 20% were used as an independent test set. All classifiers used the same training and testing sets.

2.2. Feature Extraction Methods

The feature extraction methods listed in table 1 were used with the following *standard* parameters:

- Analysis-Window 16ms
- Overlap 8ms
- Hamming-Window

[1] The entire speech database is freely available for scientific institutions. Please contact the author via email if you are interested in using that data.

0-7803-2074-3/94/$4.00 © 1994 IEEE

Method	Variable Parameters
MFCC	number of computed cepstral coefficients between 4 and 10
LPC	order of LPC filter between 8 and 14
PLP	model order between 4 and 10
RASTA-PLP	model order between 4 and 10

Table 1. Used feature extraction methods and parameters varied within the experiment

For each feature extraction method exactly one parameter (table 1) was varied within the experiment described below, all other parameters remainend constant.

2.3. Classification Methods

All computed feature vectors were classified by the following 2 connectionist and 1 statistical classifier:

- **Multilayer Perceptron (MLP):** A fully connected 3 layer perceptron was used with 10 output nodes corresponding to 10 digits. The entire input sequence was transformend to static-like vectors that are presented to the MLP. Contrary to Peeling [8], a simple trace-segmentation was used in order to get an input vector of fixed dimension[2]. The number of input nodes was dependent on the vector size of the feature extraction method used. The number of hidden units was tested from 10 to 100. The MLP was trained using the *Conjugate-Gradient Descent* optimization, included in the OGI-speechtools [9]. The training stopped when more than 200 iterations were reached, or the gradient fell below a specified value.
- **Learning Vector Quantization (LVQ):** The *Learning Vector Quantization* as described in [10] can be used to classify vectors of constant dimension, and it is used within this study as another connectionist classifier. The LVQ was performend using the optimized-learning-rate LVQ1 (OLVQ1) which is part of the *LVQ-PAK*-software package [11]. The number of reference vectors within the codebook was varied from 200 to 1000. Initially the same number of vectors was distributed over all clusters. After the end of the initialization the vectors were redistributed according to their deviations. The OLVQ1-algorithm was then iterated using 40 times the total number of used vectors.
- **Hidden Markov Modell (HMM):** A standard HMM-recognizer based on the HMM-Toolkit HTK was used [12]. The recognizer consists of 5 to 10 state word-models with output probability distributions based on N-Gaussian diagonal covariance matrices. The number of mixtures was set to 3, because a number of tests showed that the recognition performance did not increase that much when using more mixtures per state. Contrary to the above described MLP- and LVQ-classifiers, the HMM-classifier was trained and tested with *untraced* feature vectors.

3. EXPERIMENTAL RESULTS

Experiments were carried out in the following order:

```
• for every feature extraction method
    – for every model (filter) order
        ∗ train and test HMM's using 5 to 10 states
        ∗ train and test MLP's using 10 to 100 hidden nodes
        ∗ train and test LVQ's using 200 to 1000 reference vectors
    – end for
• end for
```

This experimental procedure led to a rather large amount of data. Interesting parts of that data are described below. The most important question was which feature extraction methods lead to the best recognition results. As shown in figure 1[3], RASTA-PLP and LPC clearly outperformend PLP and MFCC - regardless of which classifier was used.

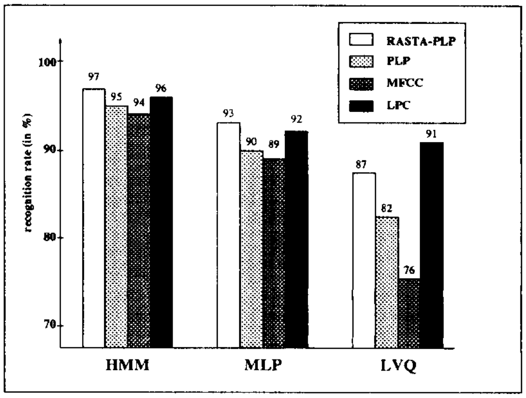

Figure 1. Best results of all combinations of feature and classification methods

In figure 1, the best results of all combinations of feature extraction and classification methods are shown. It can be seen that the different classifiers led to quite similar results. Both HMM and MLP worked best with RASTA-PLP followed by LPC. Looking at the LVQ the same feature extraction methods worked best, only the sequence changed. All classifiers showed the lowest performance when using MFCC. PLP was outperformend by LPC[4].

The following figures show the detailed performances of the 3 classifiers using RASTA-PLP or LPC. Figure 2 contains the recognition results using RASTA-PLP dependent on the order of the PLP-model used. When using HMM or MLP the best performance is achieved using RASTA-PLP with model order 7, while LVQ needs a model order

[2] Exactly the same procedure was used when applying the LVQ described below.

[3] All values in figure 1 are rounded in order to be better readable. For exact values please look below.

[4] Hermansky described the opposite in his original article about PLP while using an DTW-classifier [4].

of 9. It should be noted that the model order did not have that much influence on the HMM (with changing number of states used).

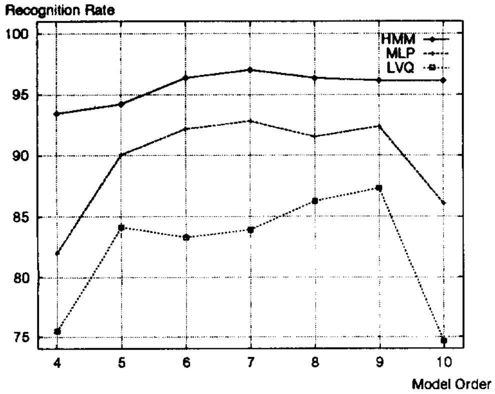

Figure 2. Recognition results using RASTA-PLP

In figure 3 the recognition results are shown using LPC, where the performance depends on the order of the LPC-filter. The model order 10 yielded the best performance using MLP and LVQ, while the HMM worked best using a model order of 11.

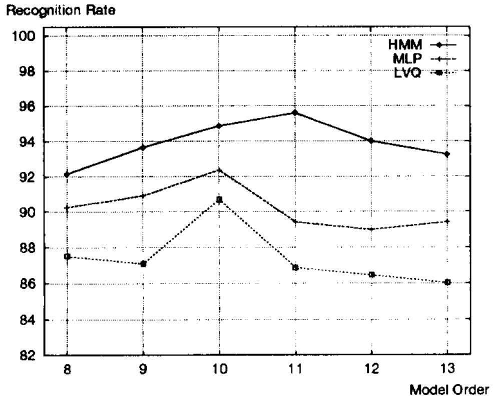

Figure 3. Recognition results using LPC

In order to get a more complete impression of the dependence of the parameters for each classifier, now only the respective best feature extraction method was chosen and the performance of that classifier is plotted against the parameter varied within the experiment.

Let's begin with the combination of HMM and RastaPLP model order 7, that provided the best performance within that experiment: **97.00%**. This result is comparable to Canavesio in [2], who achieved 98% recignition rate in a similar task. Figure 4 shows that the number of states used in the HMM is quite critical and the maximum was reached with 9 states[5].

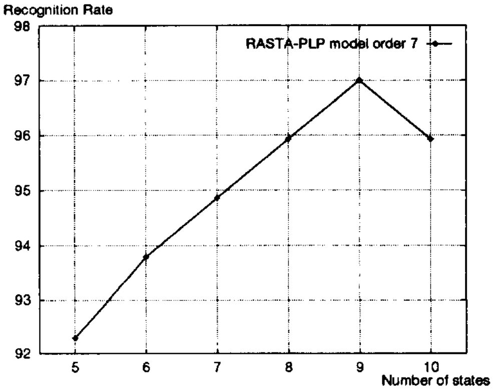

Figure 4. Plot of recognition rate of the HMM as a function of the number of states using RASTA-PLP model order 7

Using the MLP-classifier, the best result of **92.81%** was reached with RastaPLP model order 7 also. In figure 5, it can be seen that 55 hidden nodes in the MLP led to this value.

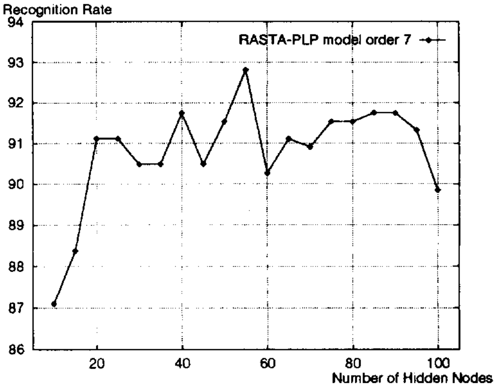

Figure 5. Plot of recognition rate of the MLP as a function of the number of hidden nodes using RASTA-PLP model order 7

[5] Tests with more than 10 states did not lead to better results and were not repeated for every possible feature extraction method.

Unlike the HMM- and MLP-results the LVQ worked best with LPC model order 10. The best result of **90.68%** was obtainend with 550 reference vectors (figure 6).

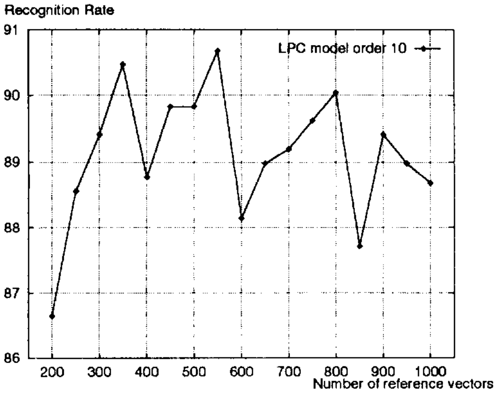

Figure 6. Plot of recognition rate of the LVQ as a function of the number of reference vectors using LPC 10th order

4. DISCUSSION

The paper describes an experiment where a number of feature extraction methods (MFCC, LPC, PLP and RASTA-PLP) were tested with 3 classification methods (HMM, MLP and LVQ). Using the recorded telephone speech database RASTA-PLP with model order 7 worked best when using either HMM or MLP and led to 97.00% and 92.81% correct recognition rates, respectively. When using LVQ, the best result of 90.68% was reached using LPC with model order 10. Especially when comparing the HMM and the MLP quite similar (relative) results are obtained. Only when using the LVQ the reached recognition rates differ not only in their obsolute but in their relative values too.

The performance gap between RASTA-PLP and LPC was not that significant as described in [5], but the model order used in RASTA-PLP is lower than in LPC.

Even though not all classifiers led to the same relative results (e.g, same sequence of feature extraction methods) it can be said that the *quality* of the different feature extraction methods is not so much dependent on the classifier used.

In future work, only the HMM will be applied and the combination of several feature vector streams will be extensively tested.

5. ACKNOWLEDGEMENTS

This study was possible thanks to the work of Hynek Hermansky, Etienne Barnard and Johan Schalkwyk from OGI (Portland), Nelson Morgan from ICSI (Berkeley), Teuvo Kohonen from HUT (Helsinki), Steve Young and Phil Woodland from CUED (Cambridge), who all wrote very usefull software packages that allowed the author to deal with the described task within a reasonable time!

REFERENCES

[1] J.Song and A.Samouelian, "A Robust Speaker-Independent Isolated Word HMM Recognizer for Operation over the Telephone Network," *Speech Communication*, vol. 13, pp. 287–295, 1993.

[2] F.Canavesio *et al.*, "HMM Modelling in the Public Telephone Network Enviroment: Experiments and Results," in *Proceedings of European Conference on Speech Technology*, (Genova, Italy), pp. 731–734, 24–26 September 1991.

[3] D.L.Thomson *et al.*, "Automatic Speech Recognition in the Spanish Telephone Network," in *Proceedings of European Conference on Speech Technology*, (Genova, Italy), pp. 957–960, 24–26 September 1991.

[4] H.Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Accoustical Society of America*, vol. 87, pp. 1738–1752, May 1990.

[5] H.Hermansky *et al.*, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in *Proceedings of European Conference on Speech Technology*, (Genova, Italy), pp. 1367–1370, 24–26 September 1991.

[6] H.Hermansky *et al.*, "RASTA-PLP Speech Analysis." International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, TR-91-069, 1991.

[7] Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals," *Speech communication*, vol. 8, pp. 45–60, March 1989.

[8] S.M.Peeling and R.K.Moore, "Isolated Digit Recognition Experiments Using the Multi-Layer Perceptron," *Speech Communication*, vol. 7, pp. 403–409, December 1988.

[9] E.Barnard and R.A.Cole, "A neural-net training program based on conjugate-gradient optimization." Technical Report CSE 89-014, Department of Computer Science, Oregon Graduate Institute of Sceince and Technology, 1989.

[10] T.Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, pp. 1464–1480, September 1990.

[11] T.Kohonen *et al.*, "LVQ-PAK - The Learning Vector Quantization Progrtam Package." Manual V2.1, Helsinki University of Technology, Finland, 1992.

[12] S.J.Young and P.C.Woodland, "HTK: Hidden Markov Toolkit V1.5." Reference Manual, Cambridge University Engineering Department and Entropics Inc., 1993.