COMPARING DIFFERENT FEATURE EXTRACTION METHODS FOR TELEPHONE SPEECH RECOGNITION BASED ON HMM's

T. Schürer

Institut für Fernmeldetechnik
Techische Universität Berlin
tilo@cs.tu-berlin.de

ABSTRACT - Robust speech recognition over telephone lines severely depends on the choice of the feature extraction method used. In the last years, several researchers made experiments with new feature extraction methods based on the *Perceptual Linear Predictive Analysis* (PLP), and showed that these methods sometimes outperform *conventional* methods like *Linear Predictive Cepstral Coefficients* (LPC) and *Mel-Frequency Cepstrum Coefficients* (MFCC).

The aim of the study described is to find the best feature extraction method for a speech recognition system running in the public switched telephone network. Because previous work showed that HMM's clearly outperformed other classification methods, continuous density HMM's are used in that study. The above mentioned and commonly used feature extraction methods (MFCC, LPC) are compared to some PLP-based methods (simple PLP and RastaPLP). Important parameters of each feature extraction method (e.g. model or filter order, number of coefficients) were modified within reasonable ranges, and the recognition performance of each computed list of feature vectors was tested with a set of different HMM's. In order to find the *optimal* HMM covering the specific data, the number of states and mixtures per state were also modified within the tests.

INTRODUCTION

State of the art HMM recognizers for telephone speech use LPC- or MFCC-analysis [Song,J. et.al., 1993], [Canevesio,F. et.al., 1991], [Thomson,D.L. et.al., 1991]. In 1990, Hermansky first introduced a new feature extraction method, called *Perceptual Linear Predictive (PLP)* - analysis [Hermansky,H., 1990]. 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 [Hermansky,H. et.al., 1991a], [Hermansky,H. et.al., 1991b]. This method is called *RelAtive SpecTrAl* (RASTA)-PLP. In order to find the best feature extraction method for a speech recognition system running in the telephone network different feature extraction methods (MFCC, LPC, PLP, RASTA-PLP) were compared. As classifier a standard continous density HMM based on the HTK-toolkit [Young,S.J. et.al., 1993] was used.

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 300 speakers was recorded. All speech data was automatically endpointed using an energy-based method after Savoji [Savoji,M.H., 1989]. Randomly selected 80% of all digits formed the training set, while the remaining 20% were used as an independent test set. All reported results are based on the independent test set.

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

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

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

HIDDEN MARKOV MODEL

Earlier experiments [Schürer,T., 1994] showed that HMM's with 9 states (7 emitting states) performed best using the speech database described. The models were built using the HTK-toolkit [Young,S.J. et.al., 1993] and are left-to-right models with no skip states and diagonal covariance matrix. No tests were made to modify the number of states for each digit-model individually. In order to find an optimal set of HMM's the following parameters were varied within the tests:

- number of gaussian mixtures per state and
- usage of 1st and 2nd time derivate of feature vector modelled by different input streams.

EXPERIMENTAL RESULTS

Within the first test all above mentioned feature extraction methods (MFCC, LPC, PLP and RASTA-PLP) were used without any time derivates. Figure 1 shows the recognition performance based on each feature extraction method versus the model/parameter order. Within this test 9-state-HMM's with 1 stream and 3 gaussian mixtures per state were applied and only the speech parts found by the speech endpoint detection were used. Figure 1 shows clearly that RASTA-PLP outperforms any other feature extraction method when using a model order of 7, the following test therefore applied exactly that combination. To avoid the explicit endpoint detection, *one* extra HMM was trained to model noise, silence, channel distortion and breath sounds. In order to find out how many gaussian mixtures per states lead to the best recognition performance, the number of mixtures per states was variied between 1 and 20. In figure 2 it can be seen that 7 gaussian mixtures per states are sufficient when using the RASTA-PLP-vectors without any time derivates. Adding the 1st time derivate led to a dramatic recognition performance improval, the performance maximum there was reached with 10 gaussian mixtures while adding the 1st and the 2nd time derivates decreased the performance.

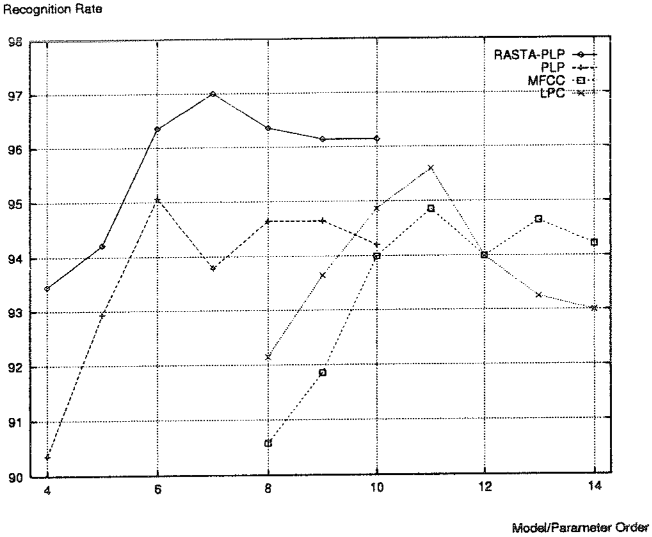

Figure 1: Plot of recognition performance versus model/parameter order of different feature extraction methods

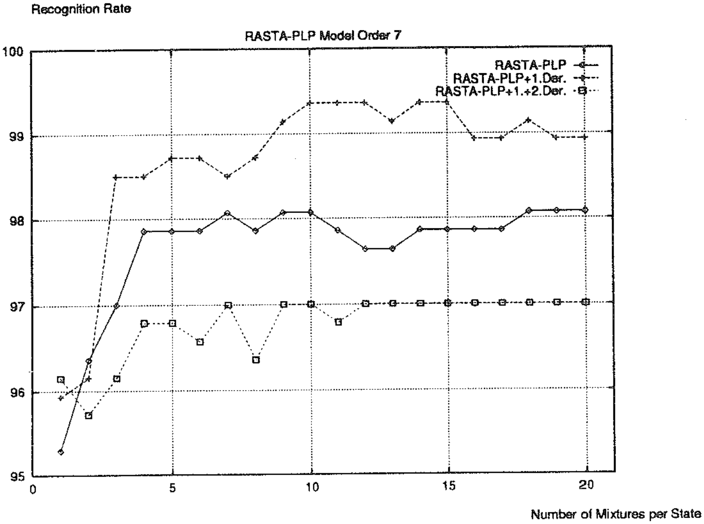

Figure 2: Plot of recognition performance using RASTA-PLP model order 7 versus number of mixtures per state

DISCUSSION

The paper describes a comparison of different feature extraction methods for telephone speech using HMM's as classifiers. It shows that RASTA-PLP (*RelAtive SpecTrAl*-PLP) clearly outperforms any other feature extraction method used in the test. Adding the 1st time derivate of the RASTA-PLP-vectors and increasing the numbers of gaussian mixtures per HMM-state to 10 led to the recognition performance of **99.36%** using the telephone speech database decribed above. Future work will be focused on the application of RASTA-PLP to connected digit recognition over the telephone network.

References

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

[Hermansky,H., 1990] Hermansky,H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Accoustical Society of America*, 87(2):1738–1752.

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

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

[Savoji,M.H., 1989] Savoji,M.H. (1989). A Robust Algorithm for Accurate Endpointing of Speech Signals. *Speech communication*, 8(1):45–60.

[Schürer,T., 1994] Schürer,T. (1994). Optimierung eines Spracherkenners für Telefonsprache. In *Elektronische Sprachsignalverarbeitung*, Berlin.

[Song,J. et.al., 1993] Song,J. et.al. (1993). A Robust Speaker-Independent Isolated Word HMM Recognizer for Operation over the Telephone Network. *Speech Communication*, 13:287–295.

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

[Young,S.J. et.al., 1993] Young,S.J. et.al. (1993). HTK: Hidden Markov Toolkit V1.5. Reference Manual, Cambridge University Engineering Department and Entropics Inc.

PITCH ESTIMATION USING DISCRETE ANALYTIC SIGNALS

T. Matsuoka, N. Hayakawa, Y. Yashiba, Y. Ishida, T. Honda and Y. Ogawa

Department of Electronics and Communication
Meiji University

ABSTRACT - This paper proposes a new method for estimating the fundamental frequency of speech signal which uses the low-pass filter and Hilbert transformer with approximately ideal frequency responses.

INTRODUCTION

As it is well known, the discrete Hilbert transform can be used for calculating discrete analytic signals. However, the Hilbert transformer with ideal frequency responses is not physically realizable because it is non-causal. In order to approximately implement such a Hilbert transformer, Schüssler[1] and Ishida[2] proposed the design method using time reversal techniques. This method first divides the impulse response of the ideal Hilbert transformer into two parts, i.e., causal and non-causal parts and then approximately realizes it by the cascade connection of the causal and non-causal filters using time reversal techniques.

In this paper, we propose a new method of estimating the fundamental frequency using the Hilbert transformer and the low-pass filter based on the method described above. Experimental results show that our method is effective to estimate the fundamental frequency of speech signal.

DESIGN OF THE HILBERT TRANSFORMER BASED ON TIME REVERSAL TECHNIQUES[1],[2]

A Hilbert transformer is a linear time-invariant system whose ideal frequency response is defined as

$$H(e^{j\omega}) = \begin{cases} -j & 0 < \omega < \pi \\ 0 & \omega = 0,\ \pi \\ +j & -\pi < \omega < 0\ . \end{cases} \qquad \cdots(1)$$

The corresponding impulse response is given by

$$h(n) = \begin{cases} \dfrac{2\sin^2(\pi n/2)}{\pi n} & n \neq 0 \\ 0 & n = 0\ . \end{cases} \qquad \cdots(2)$$

Since $h(n) \neq 0$ for $n < 0$ (for n odd), an ideal Hilbert transformer is not causal and physically realizable. In order to approximately realize such a filter, we first divide the impulse response into two parts, i.e.,

$$h(n) = h_p(n) + h_m(n) \qquad \cdots(3)$$

where $h_p(n)$ is causal and $h_m(n)$ is not causal. Using the z-transform, we get

$$H_p(z) = -H_m(z^{-1})\ . \qquad \cdots(4)$$

Considering causality, the Hilbert transformer can be block-diagrammed as Figure 1. The subfilter $H_p(z)$ is a realizable filter and the boxes labeled TIME REVERSAL have input-output relations of the form

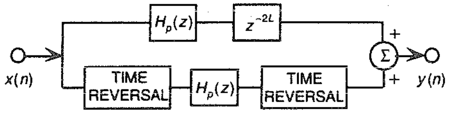

Figure 1. Block diagram of the Hilbert transformer

$$z(n) = w(-n) \qquad \cdots(5)$$

where $w(n)$ is the input sequence and $z(n)$ is the output sequence.

PITCH EXTRACTION BY A LOW-PASS FILTER AND A HILBERT TRANSFORMER

Figure 2 shows the pitch extraction system using a low-pass filter and a Hilbert transformer with approximately ideal frequency responses. In this system, the speech signal is sampled at 10kHz by using a 12 bits A/D converter, and then the sampling rate is reduced to 2 kHz by a decimation process. The decimated output is segmented in frames of 32 ms and is filtered by the low-pass filter, which can be designed by the same method as that used for the Hilbert transformer. The cut-off frequency of LPF is automatically tuned with neural networks so as to follow the fundamental frequency of speaker. The filtered output is then transferred to the Hilbert transformer.

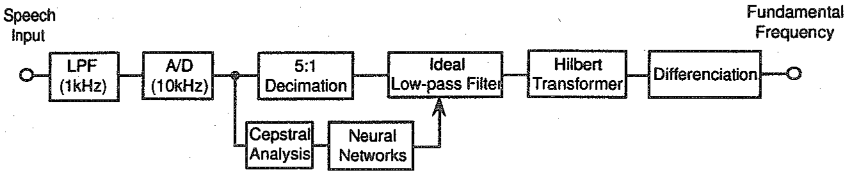

Figure 2. Structure of the extraction system for the fundamental frequency

The Hilbert transformer can be used for calculating instantaneous attributes of a time series, in particular the amplitude and frequency[3]. The instantaneous amplitude is the amplitude of the complex Hilbert transform. On the other hand, the instantaneous frequency is the time rate of change of the instantaneous phase angle. These two signals are obtained from the analytic signal. The analytic signal $a_x(t)$ of a real signal $x(t)$ is defined as:

$$a_x(t) = x(t) + jh_x(t)\ . \qquad \cdots(6)$$

Being $h_x(t)$ the Hilbert transform of $x(t)$

$$h_x(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(t')}{(t-t')}dt'\ . \qquad \cdots(7)$$

Then, the analytic signal can be expressed in modulus-argument form:

$$a_x(t) = e_x(t)\exp(j\phi_x(t))\ . \qquad \cdots(8)$$

The amplitude and instantaneous phase of the signal $x(t)$ are