METHODS FOR IMPROVED SPEECH RECOGNITION OVER TELEPHONE LINES

Alfred Hauenstein and Erwin Marschall
Siemens AG,
Corporate Research and Development, Dept. ZFE T SN 53,
81730 Munich, Germany
email: [Alfred.Hauenstein,Erwin.Marschall]@zfe.siemens.de

ABSTRACT

Robust modelling and fast adaptation to changes in transmission channels has yielded significant improvements in speech recognition over telephone lines. Robust modelling is achieved by using a special version of the LDA-transformation including a two frame context and subtraction of the mean channel seen in training. A fast maximum likelihood channel adaptation copes with variations in characteristics of transmission channel and speaker during real world operation. Evaluation of these techniques on different databases demonstrates reductions of word error rates up to 70%, suggesting that significant improvements in recognition performance may be achieved by better acoustic-phonetic modelling and fast adaptation.

1 INTRODUCTION

Speech recognition on data transmitted over real world telephone lines imposes strong requirements on speech modelling. Problems are compounded when different communication sets or transmission methods (analog, digital, cellular) are used in training and recognition.

The main application areas we tackle are spoken commands or digits for use of advanced telecommunication services (e.g. interactive voice response). Speech controlled operation is of particular importance as touch tone operation is not generally available in German public telephone networks. We therefore implemented isolated word recognition and word spotting on small vocabularies.

In this paper we investigate two approaches in order to achieve more robust modelling and a fast adaptation to changes in the transmission channel. First we introduce Linear Discriminant Analysis (LDA), which integrates a two frame context and subtracts the mean channel seen in training. Second, we add a maximum likelihood channel adaptation, which adapts very fast to the transmission channel and/or speaker characteristics.

2 THE BASELINE RECOGNIZER

Our baseline speech recognizer implements Continuous Density Hidden Markov Models (CD-HMM) with Laplacian density functions. For each frame (10 ms spaced) we extract a 51 element feature vector. It consists of 24 melscaled cepstral[1], 12 Δcepstral, 12 ΔΔcepstral, 1 energy, 1 Δenergy, and 1 ΔΔenergy components [1]. We use context-dependent diphone models. Each phoneme consists of 3 segments; each segment is modelled by 2 states with tied emission probabilities.

3 MODELLING IMPROVEMENTS

3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a well known preprocessing step for calculating a proper set of discriminant features [5], [6]. The basic idea is to find a linear transformation such that a suitable criterion of class separability is maximized. Usually the transformation is obtained as the eigenvector decomposition of the product of two scatter or covariance matrices, the total-scatter matrix and the inverse of the average within-class scatter matrix. This yields a single class-independent transformation matrix (LDA-matrix) [2].

We apply a modification of the general framework in calculating the LDA-matrix stepwise and subtracting the total mean of the original feature vectors. This avoids some numerical problems (e.g. by a pinned component of the feature vector) and guarantees robustness.

In the first step the within-class scatter matrix $\underline{S}_w$ is diagonalized yielding the eigenvector matrix $\underline{U}$ and the eigenvalue matrix $\underline{\Lambda} = \mathrm{diag}\ \{\lambda_i\}$ (weighted mean of vari-

[1] For conveniance we write 'cepstral coefficients' instead of correctly 'cepstrally smoothed spectral coefficients'.

0-7803-2431-5/95 $4.00 © 1995 IEEE

ances over classes). Let $\lambda_{max} = max_i\{\lambda_i\}$, a lower boundary $\lambda_{min} = 0.1\ \lambda_{max}$ is applied:

$$\lambda_i = \begin{cases} \lambda_i & \text{for} \quad \lambda_i \geq \lambda_{min} \\ \lambda_{min} & \text{for} \quad \lambda_i < \lambda_{min} \end{cases} \tag{1}$$

(modified whitening transformation). This yields a transformation matrix

$$\underline{B} = \underline{U} \cdot \Delta^{-1/2}. \tag{2}$$

In the second step the transformed average between-classes scatter matrix $\underline{S}_b = \underline{B}^t \underline{S}_b \underline{B}$ is diagonalized yielding eigenvectors $\underline{V}$ and eigenvalues $\{\sigma_i\}$. Let the eigenvalues be ordered according to magnitude. Neglection of the smallest eigenvalues reduces the number of components of the transformed feature vector and thus gives a systematic way of parameter reduction. In summary, given a cepstrally smoothed input feature vector c the following transformation is applied:

$$y = \underline{A}^t (c - m) \tag{3}$$

where $\underline{A} = \underline{U}\Delta^{-1/2}\underline{V}$ is the (rectangular) LDA-matrix described above. m is the total mean of all original feature vectors seen in training. Therefore m subtracts the "mean channel" and "mean speaker" as seen in training.

After transformation the mean variance of each feature component is 1. This offers the possibility to rescale each component to a given number interval.

Within this framework several decisions have to be made:

- choice of the proper classes to be discriminated
- using a single feature vector or several successive frames jointly ("supervector").
- number of components to be retained.
- handling the class "silence" in calculating $\underline{S}_w$, $\underline{S}_b$ and m in order to avoid domination of these three quantities by the "silence" class.

As our baseline system relies on the use of phoneme segments we choose these segments to be the classes being discriminated. Beside other reasons, consecutive feature vectors are correlated because we use overlapping frames. The use of supervectors thus allows to model the joint density of the observed speech more accurately. Therefore we decided to use n-frame supervectors. Results are given for the 2-frame case resulting in a 102 element supervector and retaining 51 components after transformation (so the number of parameters to be estimated is about the same as for the baseline system). With respect to the "silence" class different possibilities were tried. In the results presented the "silence" class is neglected entirely.

We apply a two-step training. Starting with a non-LDA CD-HMM the segmentation of training data with respect to class labels is determined for calculating the scatter matrices $\underline{S}_w$ and $\underline{S}_b$. The described operations deliver the LDA-matrix. In the second step an iterative training is done using LDA transformed feature vectors.

3.2 Maximum Likelihood channel adaptation

A problem of real world application of speech recognition arises from the fact, that recording conditions of training data and operating conditions differ due to variations in transmission channels, recording equipment, and human factors (stress, casual talkers, ...)

It is well known that subtracting a long-term average spectrum from the speech signal improves recognition because it estimates speaker- and/or transmission-channel properties [3], [4], [7]. Unfortunately, estimating a long-term average is not reliable in our applications, as speakers hold the telephone lines only for few seconds and utterances are very short. Furthermore, adaptation has to be performed online, in order to react immediately to user commands. Therefore we designed a very fast adaptation algorithm.

We assume that variations are an additive distortion vector $x_k(t)$ in the cepstral domain:

$$c_k(t) = c_{k0}(t) + x_k(t), \tag{4}$$

where $c_{k0}(t)$ is the feature vector seen in the training and $c_k(t)$ is the resulting distorted feature vector, with $k \in [1, ..., 24]$ denoting the component index of the feature vector[2]. The distortion vector $x_k(t)$ is modelled as normally distributed:

$$p(x_{k0}) = \frac{1}{\sqrt{2\pi} \cdot \bar{\sigma}_k} \cdot exp\left(\frac{(x_{k0} - \bar{x}_k)^2}{\bar{\sigma}_k^2}\right). \tag{5}$$

$\bar{x}_k$ and $\bar{\sigma}_k^2$ are the component dependent mean and variance of the normal distribution and are derived from the training database.

In order to determine the actual distortion during recognition we calculate a maximum likelihood (ML) estimate for each component:

[2] For adaptation we deal with the base cepstrally smoothed coefficients only, as the energy, Δ and ΔΔ components are linear superpositions of the base components.

$$\hat{x}_{k0}(t) = \frac{1}{1+\alpha_k(t)}(\alpha_k(t)\bar{x}_k + \bar{x}_k(t))\,. \qquad (6)$$

$\bar{x}_k(t)$ is the recursively computed mean of the test data. $\alpha_k(t)$ can be interpreted as a component and time (!) dependent first order filter coefficient and is calculated as follows:

$$\alpha_k(t) = \alpha_k(t_{act}) = \frac{\sigma_k^2(t)}{t_{act}\cdot\overline{\sigma}_k^2} \qquad (7)$$

t_{act} is the actually used "memory length" of the recursion:

$$t_{act} = min\{t-t_o, t_{max}\} + \Delta T, \qquad (8)$$

where t_0 is the starting frame for estimation. This means, that in order to get a fast adaptation to the current distortion we consider the last t_{act} frames only. ΔT is a positive number determining adaptation near t_0. $\sigma_k^2(t)$ is the variance of the test data. Due to algorithm simplifications and to make the computations more robust and less time consuming we set:

$$\sigma_k^2(t) = \overline{\sigma_{k0}^2}, \qquad (9)$$

where $\overline{\sigma_{k0}^2}$ is the mean of the variances of the different channels of the training data. As can be seen from eq. (7) $\alpha_k(t)$ starts with high values, because the denominator becomes small for short times t_{act} and the first term in the numerator of eq. (6) ($\bar{x}_k$) dominates the estimation. With increasing t_{act} the filter coefficient decreases and $\bar{x}_k(t)$ (the mean of the test data) becomes more important. (We apply t_{act} for the calculation of $\bar{x}_k(t)$ too: $\bar{x}_k(t) = \bar{x}_k(t_{act})$.)

We obtained best results when setting t_{max} to an upper bound of 25 frames. This means that we use, independent of the length of the utterance, at most the last 250 ms data for the ML estimation. This guarantees a fast adaptation to the most recent acoustic events.

The estimated distortion vector components $\hat{x}_{k0}(t)$ are subtracted from the extracted (distorted) feature vector components $c_k(t)$ in order to get the adapted cepstrally smoothed vector components $c_{k0}(t)$:

$$c_{k0}(t) = c_k(t) - \hat{x}_{k0}(t) \qquad (10)$$

$c_{k0}(t)$ are assumed to be the not distorted feature vector components and are used as input for the LDA.

We do not differentiate between speech and non-speech signals for the estimation. Therefore our adaptation takes into account speaker as well as transmission channel variations. In our experiments different sets of estimation parameters for speech and non-speech signals led to worse results than the "pooled" parameter set. We suppose this is due to uncertainties in the speech / non-speech discrimination algorithm.

It is important to notice that we reinitialize the estimation when a new call is set up, i.e. when a new speaker or transmission channel has to be adapted.

4 RECOGNITION EXPERIMENTS

All recognition experiments using LDA and ML channel adaptation (ML-CA) are carried out on isolated words. The vocabulary consists of the 11 German digits (including two representations for "2": '*zwo*' and '*zwei*'). The sampling rate of all data used is 8 kHz.

4.1 Databases for training and test

For training we use two different databases collected over the German public switched telephone network (PSTN). The databases used are characterized in table 1.

For training database "TIL" 193 speakers uttered approximately 30 different connected digit strings (approximately 18 000 digits in total). Database "VM" contains isolated words and is divided in one training and one test set (test and training sets are non-overlapping). The set used for training contains 591 speakers and has approximately the same size (in seconds of speech) as TIL.

name of database	properties	No. of utterances
VM	isolated words (digits and commands) recorded over the German PSTN (analog, digital and cellular); testset contains digits only	Test: 1235 Training: 15804
TIL	connected digits; same recording equipment and conditions as VM	6206
FF	isolated digits; high quality office recordings	2200
ISDN	isolated digits; recordings over the German public telephone network (analog and digital) with an ISDN telephone	2260
TESDA	isolated digits; recordings of partly corrupted utterances	2569

Tab. 1: Characterization of databases used

In all test databases the transmission channel varies from utterance to utterance. As each test utterance is a single word, we could test the speed of our adaptation algorithm. For further characterization of the test databases refer to table 1.

4.2 Recognition results

The recognition results using the TIL database for training are shown in table 2. One has to take into account that the training database contains connected digits, while the test databases consists of isolated digits only. The error rates are the sum of substitutions, insertions, and deletions.

Training database = TIL	FF	VM	ISDN	TESDA
baseline recognizer	3.0 %	9.1 %	17.5 %	11.2 %
baseline + LDA	1.4 %	6.2 %	5.0 %	9.0 %
baseline + LDA + ML-CA	1.5 %	6.1 %	4.1 %	7.0 %

Tab. 2: Error rates for different test databases using training database TIL

The recognition results for the training database VM are shown in table 3. Absolute error rates on all databases are the same or better than when using training database TIL. We attribute this result to the fact that coarticulation effects in database TIL influence the density functions of the HMM. Relative improvements are comparable to those determined using training database TIL.

Training database = VM	FF	VM	ISDN	TESDA
baseline recognizer	3.1 %	9.0 %	13.4 %	9.4 %
baseline + LDA	1.5 %	6.3 %	4.9 %	7.1 %
baseline + LDA + ML-CA	1.5 %	4.7 %	3.6 %	5.7 %

Tab. 3: Error rates for different test databases using training database VM

5 CONCLUSION AND FUTURE WORK

The results show remarkable improvements (up to 60%) in error rates on all databases when using the LDA. The LDA applied here shows better performance than the „standard" LDA. This can be attributed to the use of additional frame context in the feature vector (supervector) and to the subtraction of the total mean over the training database.

The maximum likelihood channel adaptation shows additional improvements (up to 27%). It is important to notice that the error rate did not increase but decrease when the recording equipment was the same in training and test (VM database). When the tests already showed very good results without ML-CA for high quality recordings (FF database) the error rate did not increase either.

In order to further improve the ML channel adaptation we will try to implement a more precise speech / non-speech discrimination and to improve the calculation of the parameter sets.

First experiments on 16 kHz sampled continuous speech data showed promising reductions of the error rate, when using a 3-frame supervector for LDA and a ML channel adaptation over a longer estimation interval (some seconds).

6 LITERATURE

[1] K. Zünkler, "An ISDN Speech Server Based on Speaker Independent Continuous Hidden Markov Models", in Proc. NATO ASI Cetraro, 1990.

[2] P.A. Devijver, Pattern Recognition: A Statistical Approach, Prentice Hall, 1982

[3] S. Furui, Digital Speech Processing, Synthesis, and Recognition, New York, Dekker, 1989

[4] M. Wittman et al., "Online Channel Compensation for Robust Speech Recognition", Proc. Eurospeech 1993, pp. 1251-1254.

[5] P.F. Brown, The Acoustic-Modelling Problem in Automatic Speech Recognition, PhD thesis, CMU, Pittsburgh, Pa, 1987

[6] G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition", Proc. ICASSP, pp.556-559, Glasgow, Scotland, May 1989

[7] Yeunung Chen, "Cepstral Domain Stress Compensation for Robust Speech Recognition"; Proc. ICASSP, pp 717-720, Dallas, USA, April 1987.

ACKNOWDLEDGEMENTS

The authors want to thank Dr. Harald Höge for initiating our work on the topics mentioned and for submitting many fruitful ideas during the whole process of implementation and test.