CONTINUOUS MIXTURE DENSITIES AND LINEAR DISCRIMINANT ANALYSIS FOR IMPROVED CONTEXT-DEPENDENT ACOUSTIC MODELS

X. Aubert, R. Haeb-Umbach, H.Ney

Philips GmbH Research Laboratories Aachen, P.O. Box 1980, D-5100 Aachen, Germany
aubert@pfa.philips.de

ABSTRACT

The aim of this paper is threefold:

- To extend the Linear Discriminant Analysis (LDA) experiments presented at the ICASSP-92 to context-dependent models and speaker-independent large vocabulary continuous speech recognition.
- To compare two variants of using mixture densities: the state-specific modeling and the "monophone-tying" approach where densities are shared across the states relevant to the same phoneme.
- To present results on the DARPA RM task for both speaker-dependent and speaker-independent parts.

Using triphone models based on LDA and continuous mixture densities, significant improvements have been observed and the following word error rates have been achieved: for the speaker-dependent (SD) part, 7.8 % without grammar and 1.5 % with word pair and for the speaker-independent (SI) part, 17.2 % and 4.6 % respectively. These scores are averaged over 1200 SD or SI evaluation sentences and are among the best published so far, on the RM database.

1. INTRODUCTION

This paper describes some recent progress of the Philips large-vocabulary continuous-speech recognition system that has been adapted to and tested on the DARPA Resource Management (RM) task [1]. The characteristic features of the baseline system [2] are :

- A large-sized acoustic vector which includes first and second-order time derivatives is used. This "slope-curvature" vector is not split into separate streams.
- The acoustic-phonetic modeling relies upon continuous mixtures of Laplacian densities.
- Viterbi criterion is used in training and recognition.
- Decoding proceeds time-synchronously in one pass.

Significant improvements have been achieved by combining linear discriminant analysis and triphone models, the word error rate being reduced by one fifth or more. For the speaker-independent application, this result has been obtained with a single linear transformation. To reach the best performance, a large number of mixture density components is needed. Therefore, two variants of continuous mixture densities have been investigated, namely, the state-specific (non-tied) modeling and the "monophone tying" approach where all triphone states relevant to the same central phoneme share a common subset of densities.
In the following sections, the main components of the system are described and experimental results are presented for both SD and SI parts of the RM task.

2. LINEAR DISCRIMINANT ANALYSIS

The basic idea of linear discriminant analysis (LDA) is to find a linear transformation such that the class separability is increased [3]. However, when applying LDA to speech recognition, a number of options are left open. In a previous study [4], three main levels of difficulties have been identified and given efficient solutions:

- What is the most appropriate class definition? Context-dependent phoneme states appear to be the most suitable choice among various possibilities.
- How are the speech data to be labeled in terms of these classes ? The class affiliation of each training pattern is automatically obtained as a byproduct of a standard HMM training done in the original space. A multiple-step training strategy is thus necessary [4].
- Which original acoustic features should be considered and how many components should be retained after transformation ? The inclusion of time differences appears very important and moreover, adjoining several centi-second frames prior to the transformation leads to improved results. Consequently, three successive "slope-curvature" vectors have been taken as input to the LDA. The 35 first components have been retained in the transformed space.

3. CONTEXT-DEPENDENT UNITS

First, a particular treatment has been given to the function words such as articles, prepositions etc. The lexicon has been therefore divided in two subsets, one including the most frequent short function words (SFW) and the other with all remaining words, to model separately the same contexts occurring in both lexical categories. Next, word-internal triphones have been chosen to capture the most relevant contextual influences. No across-word triphones have been considered, word beginnings and endings being treated as non-specific contexts.

0-7803-0946-4/93 $3.00

For the present application i.e. the RM task, the lexicon (very similar to the one used in [5]) contains 991 entries and there are 46 base phoneme-like units not including silence. Triphones have been selected on the basis of their number of occurrences in the training script. Table 1 contains the main information about the phoneme inventory.

Table 1: Within-Word Triphone RM Inventory

SD Script :	600 Sentences, 5237 words	
	# Occur.	# Triphones
SFW Contexts	≥ 5	134
$\neq$ SFW Contexts	≥ 7	635
Total Number	46 Monophones+769 Triphones	

SI Script :	2880 Sentences, 25,228 words	
	# Occur.	# Triphones
SFW Contexts	≥ 22	164
$\neq$ SFW Contexts	≥ 25	789
Total Number	46 Monophones+953 Triphones	

Each word is represented by a linear sequence of phonemes that corresponds to its standard pronunciation. As the lexicon includes several abbreviations (like USA, C1, etc.), an optional silence has been inserted between the constitutive elements (letters or digits) of these lexical entries.

4. ACOUSTIC-PHONETIC MODELING

Each subword unit is represented by a three-state left-to-right HMM. Let $x_1, ..., x_t, ..., x_N$ be the time sequence of observation vectors and $s = 1, ..., S$ be any state.

4.1 Emission Probability Distributions

The emission probability density function associated with each state s is assumed to be of the form [6]:

$$q(x_t|s,\Theta_s) = \sum_{k=1}^{K(s)} w_{k,s}\, b_k(x_t|s,\theta_{k,s})\,, \qquad (1)$$

$$\sum_{k=1}^{K(s)} w_{k,s} = 1\,, \quad w_{k,s} \geq 0\,, \qquad (2)$$

where each mixture component $b_k(.|.)$ is a unimodal density such as a Gaussian or Laplacian with parameter vector $\theta_{k,s}$ (e.g. mean and covariance), $w_{k,s}$ are the mixture weights subjected to the stochastic constraints (2) and $K(s)$ is the number of component densities. In the current system, Laplacian-type densities have been implemented, the vector of absolute deviations being pooled over all states.

4.2 Monophone Tying

So far, it has been assumed that the component densities are specific to each HMM state. When using context-dependent units like triphones as shown in table 1, each CI phoneme gives rise (on the average) to about 50 HMM states that are treated separately. In order to reduce the number of component densities, some partial sharing of the mixture parameters called "monophone tying" has been considered [7]. All states relevant to the same monophone (central part of triphone) are constrained to share a common subset of densities and are only discriminated through their weights.

Compared to the semi-continuous HMM approach [8], where each state distribution resorts to the whole set of density components, monophone tying implies some a priori restriction. On the other hand, the weight matrix is no longer full but exhibits a sparse structure that can easily be exploited to reduce the number of non-sero elements by two orders of magnitude. It is then possible to keep a fairly large number of densities without leading to a prohibitive number of weights. This compromise between acoustic resolution and number of free parameters is harder to achieve with the semi-continuous (fully tied) technique.

5. LEARNING ALGORITHM

The parameters of the system are trained with the Viterbi approximation, e.g. using the single best state sequence. As a starting point, the training sentences are linearly segmented into phoneme states and each mixture is initialised with a single Laplacian density. In the course of the training process, new density components are gradually introduced on a data-driven way to provide an improved statistical fitting of the data [2].

Two methods inherited from classification techniques [3] have been applied for estimating the mixture parameters. The one-to-one correspondence between observation and state, as provided by the Viterbi approximation, is fully exploited. In particular, the number of observations $\hat{n}_s$ assigned to any state s, is known after each time alignment. When applied to a mixture of continuous densities (Eqs. 1 & 2), the standard maximum likelihood (ML) method leads to an implicit set of nonlinear equations that can only be solved iteratively starting from initial estimates :

$$\hat{w}_{k,s} = \sum_{x_t \to s} \frac{\hat{P}(k|x_t,s,\hat{\Theta}_s)}{\hat{n}_s}\,,\ \hat{r}_{k,s} = \frac{\sum_{x_t \to s} \hat{P}(k|x_t,s,\hat{\Theta}_s) x_t}{\sum_{x_t \to s} \hat{P}(k|x_t,s,\hat{\Theta}_s)}\,, \qquad (3)$$

where $x_t \to s$ means that observation x_t has been assigned to state s during time-alignment and $\hat{P}(k|x_t,s,\hat{\Theta}_s)$ is the ML estimate of the a posteriori probability of observation x_t belonging to density k, computed with:

$$\hat{P}(k|x_t,s,\hat{\Theta}_s) = \frac{\hat{w}_{k,s}\, b_k(x_t|s,\hat{\theta}_{k,s})}{\sum_{l=1}^{K(s)} \hat{w}_{l,s}\, b_l(x_t|s,\hat{\theta}_{l,s})} \qquad (4)$$

This solution is a particular instance of the "Expectation and Maximisation" (EM) procedure [9]. To reduce the computations, a common approximation consists in replacing the sum over the mixture components in Eq. 1 by the maximum operation:

$$q(x_t|s,\Theta_s) = \max_{k=1...K(s)} w_{k,s}\, b_k(x_t|s,\theta_{k,s}) \qquad (5)$$

This amounts to neglecting the overlap between the component distributions and allows each observation to be allocated to only one particular density k. The ML estimates of the mixture parameters are then :

$$\hat{w}_{k,s} = \frac{\hat{n}_{k,s}}{\hat{n}_s} , \quad \hat{r}_{k,s} = \frac{\sum_{x_t \to k} x_t}{\hat{n}_{k,s}} , \qquad (6)$$

where $\hat{n}_{k,s}$ is the number of observations assigned to component $k = 1, ...K(s)$ of state s in the "current" iteration.

When mixture densities are used without tying, training proceeds by combining Viterbi alignment with the approximated mixture estimation (Eqs. 5-6). Since the resulting algorithm is very economical in both memory and computational costs, it is possible to reach a very high acoustic resolution. Concerning monophone-tying, the total number of Laplacians has been deliberately limited to a few thousand. The models are initialised by means of context-independent HMMs and trained using Viterbi alignment coupled with the EM algorithm (Eqs. 3-4).

6. APPLICATION to SD RM

The Speaker-Dependent part of RM [1] comprises 12 speakers (7 Males & 5 Females). The sampling frequency is 16 kHz. All parameters have been trained from the 600 training sentences (5,237 words), including one LDA transformation for each speaker. The total number of HMM states is $3 * (769 + 46) = 2445$ plus a silence model. Three systems have been considered differing in the observation vector and the mixture density modeling:

1. NO-LDA, NON-TIED : System 1 uses a 33 component acoustic vector, without LDA. There are 12,000 densities in total, i.e. 408,000 parameters / speaker.
2. LDA, NON-TIED : System 2 differs from System 1 only in LDA and has 432,000 parameters / speaker.
3. LDA, MONOPHONE-TIED : the densities are now tied across monophones. There are 5,000 shared densities which amounts to 215,000 parameters.

The three systems have been applied to and tuned on the development sentences, either without grammar (NG) or with the official word pair (WP). The test set perplexities are respectively 991 and approximately 60. The same search parameters (beam threshold and word penalty) have been used for all speakers. Word error rates (WER) include deletions (Del), insertions (Ins) and substitutions (Sub).

Table 2: SD System Comparison on Development Set (12 Speakers, 1200 Sentences, 10,242 Words)

	NO GRAMMAR		WORD-PAIR	
SYS	DEL/INS	WER %	DEL/INS	WER %
S-1	1.8/0.95	9.2	0.40/0.3	2.0
S-2	1.3/0.90	7.2	0.35/0.2	1.6
S-3	1.2/0.75	6.8	0.30/0.2	1.3

Some comments follow on the results of Table 2:

- System 1 achieves already good performance due to the high acoustic resolution of the mixture densities.
- System 2 clearly shows the improvement gained by LDA, the word error rates being reduced by one fifth.
- The "monophone-tied" system 3 demonstrates further progress attributed to the somewhat smoother distributions provided by the EM algorithm.

The final version of the recognisers has been run on the evaluation sentences. Table 3 summarises the results of the 2 LDA-based systems and leads to the following conclusions:

- The "monophone tied" system 3 is consistently better particularly in the word-pair case.
- Compared to the results obtained by other groups, the scores achieved by System 3 are the best reported so far for the SD-part of RM.

Table 3: Evaluation of 2 SD Systems over 4 test sets (12 Speakers, 1200 Sentences, 10,288 Words)

SET	NO-GR. WER %		WORD-PAIR WER %	
	SYS-2	SYS-3	SYS-2	SYS-3
FEB89	7.4	6.9	1.6	1.45
OCT89	8.3	7.9	1.7	1.40
FEB91	8.3	7.6	2.0	1.20
SEP92	9.2	8.6	2.6	1.80
AVER.	8.3	7.8	2.0	1.50

7. APPLICATION to SI RM

All parameters have been estimated from the 3,990 training sentences pronounced by 109 speakers. The acoustic models have been trained separately for males and females. Therefore, the training set has been partitioned into male (78 speakers, 2,830 sentences) and female (31 speakers, 1,160 sentences) parts. One single LDA transformation has been used, the gender-dependent modeling concerning only the mixture parameters. During recognition, both male and female models are considered and the decoded word string with the best overall score is taken as the unique sentence hypothesis. The total number of HMM states is $3 * (953 + 46) = 2998$ plus a silence model, per gender.

The February and October 89 test sets have been used for development purpose (600 sentences, 20 speakers). An important step consisted in applying LDA in a speaker-independent mode, based on one single transformation estimated over all (M & F) training speakers. The figures in Table 4 clearly demonstrate the usefulness of LDA for speaker-independent recognition, the word error rates being reduced by one quarter or more. Moreover, the influence of the acoustic resolution has been studied either by using still more densities or by tying a reduced number of density components across monophones.

Table 4: SI Development: Effect of LDA and Acoustic Resolution (Word Error Rates are in %)

LDA?	Tying?	#densities	NG-WER	WP-WER
NO	NO	2*20000	23.6	6.5
YES	NO	2*20000	17.5	4.6
YES	NO	2*30000	17.0	4.2
YES	YES	2*7500	16.9	4.1

By increasing the number of densities from 20,000 to 30,000 for each gender, a small but consistent improvement is observed for the "non-tied" mixture system and this has been confirmed over all test sets. Concerning monophone tying, encouraging results have been achieved on the development set as similar recognition scores could be obtained with approximately 7,500 Laplacians.

For evaluation, 3 LDA-based systems have been used:

1. CI, NON-TIED: System 1 uses Context-Independent phoneme models consisting of 7,500 densities for each gender, giving a total of 540,000 parameters.
2. CD, NON-TIED: System 2 uses Context-Dependent models with approximately 30,000 density components for each gender, representing a total of 2,160,000 acoustic parameters.
3. CD, MONOPHONE-TIED: System 3 uses mixture densities tied across monophones with 7,500 component densities per gender, about 675,000 acoustic parameters including the mixture weights.

The first and third systems are based on the same number of Laplacians which allows to clearly observe the effects of tying. The results are summarized in Table 5 for the four SI evaluation sets, i.e. a total of 40 test speakers and 1200 sentences, under the same conditions as before.

Table 5: SI Evaluation over 4 Test-Sets (10,288 W.) WER are given for the NG/WP cases in %.

SYSTEM	F-O89	FEB91	SEP92	AVER.
S-1 CI	28.3/6.3	25.8/5.7	33.8/9.8	29.0/7.0
S-2 CD	17.0/4.2	15.3/4.3	19.4/6.0	17.2/4.6
S-3 CD+T	16.9/4.1	16.4/3.8	22.2/7.6	18.0/4.9

This SI evaluation leads to the following conclusions:

- The performance of System 1 is relatively low due to its coarse acoustic-phonetic resolution but provides reference values to evaluate the benefit of CD modeling and "monophone-tying".
- As opposed to the SD experiments, the best SI results are obtained with the non-tied System 2, presumably due to a lack of acoustic resolution in the monophone-tied System 3 for this SI task.
- In the No-Grammar case, both CD systems have achieved results that are among the best compared with those obtained by other research groups on the same task. System 2 in particular has produced the best performance so far, for this No-Grammar case.
- In the Word-Pair case, the results of both CD systems compare favorably with most approaches though they are not among the best reported. This could be due to the absence of across-word models.

For the September-92 DARPA RM benchmark tests, the described systems achieved the smallest word error rates in 3 over 4 cases (SD for both NG and WP and SI for NG).

Acknowledgement

The system and results described in this paper are based on a couple of projects that were carried out at Philips Research over the last 6 years and were supported by both the German Ministry of Research and Technology (BMFT grants 413-5839-ITM 8401, -8801, 01-IV-102B) and the European Community (ESPRIT Project 2104).

References

[1] P.J. Price, W.Fisher, J. Bernstein, D. Pallett, "A Database for Continuous Speech Recognition in a 1000-Word Domain", *Proc. ICASSP'88*, New-York, NY, pp 651-654, 1988.

[2] H. Ney, "Experiments on Mixture-Density Phoneme-Modelling for the Speaker-Independent 1000-Word Speech Recognition DARPA Task", *Proc. ICASSP'90*, Albuquerque NM, pp 713-716.

[3] R.O. Duda, P.E. Hart: "Pattern Classification and Scene Analysis", Wiley, New York, 1973.

[4] R. Haeb-Umbach, H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", *Proc. ICASSP'92*, San Francisco, CA, pp I(13-16), 1992.

[5] K.F.Lee, "Large Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX system", Ph.D Thesis, CMU-CS-88-148, 1988.

[6] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi, "Recognition of Isolated Digits Using HMMs with Continuous Mixture Densities", AT&T Tech. J, Vol.64, No.6, pp 1211-1234, 1985.

[7] D.B. Paul, "The Lincoln Robust Continuous Speech Recognizer", *Proc. ICASSP'89*, Glasgow, Scotland, pp 449-451, 1989.

[8] X.D. Huang, M.A. Jack, "Semi-Continuous Hidden Markov Models for Speech Recognition", Comp. Sp. and Lang., 3, pp 239-251, 1989.

[9] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Royal Statistical Soc. Ser. B (methodological), vol 39, pp 1-38, 1977.