ON THE ROBUSTNESS OF LINEAR DISCRIMINANT ANALYSIS AS A PREPROCESSING STEP FOR NOISY SPEECH RECOGNITION

Olivier Siohan

CRIN–CNRS & INRIA Lorraine
BP 239, F-54506 Vandœuvre–lès–Nancy, France

ABSTRACT

This paper addresses the problem of speech recognition in a noisy environment by finding a robust speech parametric space. The framework of Linear Discriminant Analysis (LDA) is used to derive an efficient speech parametric space for noisy speech recognition, from a classical static+dynamic MFCC space. We first show that the derived LDA space can lead to a higher discrimination than the MFCC related space, even at low signal–to–noise ratio (SNR). Then, we test the robustness of the LDA space to variations between the training and testing SNR. Experiments are performed on a continuous speech recognition task, where speech is degraded with various noises: Gaussian noise, F16, Lynx helicopter, autobus, hair dryer. It was found that LDA is highly sensitive to SNR variations for white noises (Gaussian, hair dryer), while remaining quite efficient for the others.

1. INTRODUCTION

In the past few years, researchers have focussed their attention on finding robust acoustic features effective for automatic speech recognition. Linear Discriminant Analysis (LDA) has become a popular approach to improve discrimination in a speech feature space. This led to improvements in recognition performances for both small and large vocabulary system [1, 2, 3, 4]. However, very little work has been reported to determine the robustness of LDA on a continuous noisy speech recognition task to mismatches between the training and the testing acoustical environment.

In this work, a MFCC+ΔMFCC vector space was transformed into a more discriminative space, using LDA. Our experiments on a continuous speech recognition task show the effectiveness of a parametric space obtained from LDA transformation for various noisy speech, compared to a classical MFCC space. We study the sensitivity of feature vectors obtained from a LDA transformation to variations between training and testing environmental conditions. Our results show that this sensitivity is highly noise dependent. In some noisy conditions, the recognition rate drastically drops when the training and testing signal–to–noise ratio (SNR) does not match, while on others noisy conditions, the LDA parametric space remains quite effective. Of course, at high SNR, we observe that the LDA remains efficient when training and testing SNR differs.

This paper is organised as follows. Section 2 presents the framework of LDA and its application to phoneme discrimination. Experiments and results are given in Section 3. Section 4 concludes the paper.

2. LINEAR DISCRIMINANT ANALYSIS

LDA aims at improving discrimination between classes in a vector space, by finding a linear transformation from a D-dimensional vector space to a d-dimensional vector space. A dimensionality reduction of the vector space ($d \leq D$) can optionally be performed [5]. Let $\mathbf{X}$ be a D-dimensional vector, and $\mathbf{U}$ a $D \times d$ transformation matrix. The d-dimensional transformed vector is then expressed as $\mathbf{U}^t\mathbf{X}$. The transformation is defined according to the usual criterion which maximises $\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B})$, where $\mathrm{tr}(\mathbf{m})$ denotes the trace of matrix $\mathbf{m}$. $\mathbf{W}$ and $\mathbf{B}$ are the within and between class covariance matrices defined as:

$$\mathbf{B} = \frac{1}{N}\sum_{k=1}^{K} n_k(\mu_k - \mu)(\mu_k - \mu)^t \quad (1)$$

$$\mathbf{W} = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{n_k}(x_{kn} - \mu_k)(x_{kn} - \mu_k)^t \quad (2)$$

where N denotes the total number of training patterns, K the number of classes, and n_k the number of training patterns of the kth class. The mean μ_k of the kth class and the overall mean μ are given by:

$$\mu_k = \frac{1}{n_k}\sum_{n=1}^{n_k} x_{kn} \quad (3)$$

0-7803-2431-5/95 $4.00

$$\mu = \frac{1}{N}\sum_{k=1}^{K} n_k \mu_k \quad (4)$$

where x_{kn} is the nth training pattern from the kth class.

Using the optimisation criterion previously defined, it can be shown that the d column vectors of the transformation matrix $\mathbf{U}$ are the d eigenvectors associated to the d largest eigenvalues of the matrix $\mathbf{W}^{-1}\mathbf{B}$. As $\mathbf{W}^{-1}\mathbf{B}$ is not a symetric matrix, the computation of all eigenvectors and eigenvalues is not trivial. In practice, we use the method described in [6]. Let $\mathbf{C}$ be the unitary matrix diagonalizing $\mathbf{W}$ to $\mathbf{L}$, $\mathbf{W} = \mathbf{CLC}^t$. Let $\mathbf{V}$ be the unitary matrix whose column vectors are the d eigenvectors corresponding to the d largest eigenvalues of the symmetric matrix $\mathbf{S} = \mathbf{L}^{-1/2}\mathbf{C}^t\mathbf{BCL}^{1/2}$. Then, the matrix $\mathbf{U}$ is given by $\mathbf{U} = \mathbf{CL}^{-1/2}\mathbf{V}$.

When LDA is applied as a preprocessing step in a continuous speech recognition system, we have to choose what is the best definition of classes we want to discriminate. As our continuous speech recognition system VINICS uses stochastic trajectory phoneme models [7], and is therefore a phoneme based recogniser, we decide to associate a class to a phoneme. The unique transformation matrix $\mathbf{U}$ is derived from a training corpus labelled at phonetic level.

3. EXPERIMENTS AND RESULTS

3.1. Experimental settings

Experiments were performed on a continuous speech recognition task in a speaker dependent mode, using our VINICS speech recognition system based on Stochastic Trajectory Models [7].

The database was recorded from 4 different French speakers. No speaker selection was performed to optimise the recognition rate. 79 sentences were read by each speaker as training material. Test speech text consisted of 241 sentences with 1482 words read by each speaker. Vocabulary–independent training was used: the vocabulary of the training text has been designed to have little coverage over that of recognition text. The recognition task has a vocabulary of 1011 words, with a bigram word perplexity of 25. We did not use word transition probabilities, so the effective perplexity is larger than 25. A 13^{th} order MFCC was applied on speech signals sampled at 16kHz, with a frame shift of 10ms using a window length of 25.6ms. ΔMFCC computed using a classical regression were added to the MFCC. We projected this 26-dimensional vector space into a 20-dimensional vector space. No attempt was made to optimise the dimension of the projection space.

32 context–independent phone models including one silence model were designed for all experiments. The acoustic models were trained using the EM algorithm [8]. In all experiments, we used diagonal covariance matrices.

We used five different kinds of noise: Gaussian noise, aircraft noises (F-16 and Lynx helicopter) taken from the NOISEX database [9], vehicle noise recorded from inside a moving city bus, and a hair dryer noise. Noises were subsequently added to the speech waveform at various SNR, from 0dB to 36dB, with a 6 dB step.

All experiments were done in a cross SNR mode, described as follows:

1. The LDA transformation matrix $\mathbf{U}_{\mathrm{SNR_{ref}}}$ is learnt from the labelled training corpus at a given reference SNR, called $\mathrm{SNR_{ref}}$.
2. The $\mathrm{SNR_{ref}}$ dB training corpus is transformed using $\mathbf{U}_{\mathrm{SNR_{ref}}}$ into a LDA parametric space.
3. Acoustic models are built using the transformed training corpus from step 2. These models are specific to speech at $\mathrm{SNR_{ref}}$ dB in the LDA parametric space.
4. The $\mathrm{SNR_{test}}$ dB testing corpus is transformed using $\mathbf{U}_{\mathrm{SNR_{ref}}}$ computed at step 1. The $\mathrm{SNR_{test}}$ dB training corpus is transformed too, and is used only for the phonetic evaluation of the models.
5. The testing corpus obtained from step 4 is recognised using models provided from step 3. This is the cross SNR experiment.

It should be noted that the cross SNR experiment corresponds to the practical situation where the transformation matrix and the acoustic models are derived from a given environment ($\mathrm{SNR_{ref}}$), and are used to recognise speech in a different environment. ($\mathrm{SNR_{test}}$).

3.2. Results

We performed two set of experiments. The first one consisted in evaluating the phonetic recognition rate on the training database obtained from step 4. These experiments were performed to compare the discrimination ability between acoustic models built from the transformed speech in cross SNR mode, to models derived from speech parametrised in a classical MFCC space. Of course, the vector dimensions are the same in both MFCC and LDA space: 13 coefficients in the MFCC space, and the LDA transformation projects a 26 dimensional MFCC+ΔMFCC vector into a 13 dimensional vector. Results are in term of % correctly recognised phonemes on the labelled training corpus, averaged over the 4 speakers. Results are given for 3 of the 5 kinds of noises, the Gaussian noise, the Lynx helicopter noise, and the Hair dryer noise (cf. Table 1). When the testing and training SNR are matched ($\mathrm{SNR_{test}} = \mathrm{SNR_{ref}}$),

the acoustic modelisation is clearly more effective in the LDA domain than in the MFCC domain, for every SNR and every noise.

When the training and testing SNR differ ($SNR_{test} \neq SNR_{ref}$), the results are highly noise dependent. For Gaussian noise, phonetic recognition rate drastically falls in the LDA parametric domain. For example, when the testing SNR is 6dB and the training SNR is 0dB, only 16% of the training phonemes are correctly recognised using models built in the LDA space, against 42% in the MFCC space. For the Lynx noise, it appears that LDA space results outperform those from the MFCC space. But for the Hair dryer noise at low SNR, MFCC space is more noise resistent than LDA space and vice versa at high SNR.

The second set of experiments consisted in evaluating the recognition performance on the test corpus, obtained from step 4. Experiments were performed only on speech parametrised in the LDA domain. This time, we used a 20 dimensional vector space. The results are given in term of % accuracy where for N tokens, S substitution errors, D deletion errors and I insertion errors, accuracy is expressed as $[(N - S - D - I)/N] \times 100\%$. The HTK toolkit [10] is used for scoring the recognition results, which are averaged over the 4 speakers (cf. Table 2).

Our recognition system strategy prematurely cuts branches with low probabilities in the sentence search stage, leading to very low (and biased) recognition rates, especially when acoustic models are inaccurate. This explains the meaningless results in Table 2, for Gaussian white noise when training and testing SNR does not match. Results from Table 2 are in agreement with those from Table 1. It appears that LDA parametric space is very efficient for noisy speech recognition when the training and testing SNR are identical. For example, the recognition rate at 6dB is 97.4% (averaged over all noise types) of the recognition rate at 36dB. The sensitivity to SNR variations is noise dependent, especially at low SNR. For example, when the training is performed at 18dB, recognition rates are still good when testing at 12dB with F16, Lynx and Bus noises, but fall with the Hair dryer noise (15% lower) and collapse with the Gaussian noise (43% lower). At high SNRs (>24dB), the recognition rate is quite insensitive to small variations around the reference SNR, the worse case still being with Gaussian noise.

SNR ref	SNR test — Gaussian noise							
	0dB		6dB		12dB		18dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
0dB	82	76	16	42	–	–	–	–
6dB	13	46	84	78	23	56	–	–
12dB	–	–	22	59	86	81	45	68
18dB	–	–	–	–	48	68	87	83
SNR ref	18dB		24dB		30dB		36dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
18dB	87	83	68	75	–	–	–	–
24dB	71	75	88	84	80	79	–	–
30dB	–	–	83	80	88	84	86	83
36dB	–	–	–	–	87	83	88	84
SNR ref	SNR test — Lynx noise							
	0dB		6dB		12dB		18dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
0dB	83	79	67	53	–	–	–	–
6dB	68	58	86	82	79	68	–	–
12dB	–	–	80	69	88	84	85	73
18dB	–	–	–	–	84	76	88	85
SNR ref	18dB		24dB		30dB		36dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
18dB	88	85	86	79	–	–	–	–
24dB	87	80	88	85	88	82	–	–
30dB	–	–	88	83	88	85	88	84
36dB	–	–	–	–	88	84	88	85
SNR ref	SNR test — Hair dryer noise							
	0dB		6dB		12dB		18dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
0dB	82	75	22	43	–	–	–	–
6dB	19	45	84	78	41	55	–	–
12dB	–	–	42	61	86	81	62	68
18dB	–	–	–	–	68	69	87	83
SNR ref	18dB		24dB		30dB		36dB	
	LDA	CEP	LDA	CEP	LDA	CEP	LDA	CEP
18dB	87	83	78	74	–	–	–	–
24dB	81	74	88	84	84	78	–	–
30dB	–	–	85	80	88	84	87	82
36dB	–	–	–	–	87	83	88	85

Table 1. Phonetic evaluation (%) averaged over all speakers.

4. CONCLUSION

In this paper, we first shown that a LDA transformation can lead to a parametric space giving a high accuracy to noisy speech recognition, compared to a classical MFCC space, even at very low SNR (at 6dB, the recognition rate is

SNR	SNR test — Gaussian noise						
ref	0dB	6dB	12dB	18dB	24dB	30dB	36dB
0dB	88.65	1.89	–	–	–	–	–
6dB	-1.62	95.09	7.03	–	–	–	–
12dB	–	12.28	96.58	44.59	–	–	–
18dB	–	–	54.35	96.83	80.57	–	–
24dB	–	–	–	91.78	97.35	92.04	–
30dB	–	–	–	–	97.06	97.89	97.03
36dB	–	–	–	–	–	98.55	98.23
SNR	SNR test — F16 noise						
ref	0dB	6dB	12dB	18dB	24dB	30dB	36dB
0dB	86.99	18.99	–	–	–	–	–
6dB	39.78	96.61	77.70	–	–	–	–
12dB	–	87.41	96.83	90.72	–	–	–
18dB	–	–	97.88	98.18	97.10	–	–
24dB	–	–	–	98.11	98.38	97.96	–
30dB	–	–	–	–	98.11	98.18	97.84
36dB	–	–	–	–	–	98.60	98.62
SNR	SNR test — Lynx noise						
ref	0dB	6dB	12dB	18dB	24dB	30dB	36dB
0dB	93.66	67.29	–	–	–	–	–
6dB	91.67	96.90	89.05	–	–	–	–
12dB	–	98.25	98.30	97.66	–	–	–
18dB	–	–	98.57	98.60	97.86	–	–
24dB	–	–	–	98.55	98.43	98.20	–
30dB	–	–	–	–	98.25	98.25	98.36
36dB	–	–	–	–	–	98.72	98.69
SNR	SNR test — Hair dryer noise						
ref	0dB	6dB	12dB	18dB	24dB	30dB	36dB
0dB	87.48	2.01	–	–	–	–	–
6dB	3.95	93.87	27.55	–	–	–	–
12dB	–	42.65	96.38	69.06	–	–	–
18dB	–	–	81.14	96.44	86.55	–	–
24dB	–	–	–	96.49	97.89	95.72	–
30dB	–	–	–	–	98.23	98.30	98.18
36dB	–	–	–	–	–	98.50	98.48
SNR	SNR test — Bus noise						
ref	0dB	6dB	12dB	18dB	24dB	30dB	36dB
0dB	91.31	75.44	–	–	–	–	–
6dB	85.26	97.34	93.91	–	–	–	–
12dB	–	96.25	97.87	96.86	–	–	–
18dB	–	–	98.04	98.62	98.26	–	–
24dB	–	–	–	98.60	98.68	98.43	–
30dB	–	–	–	–	98.75	98.77	98.77
36dB	–	–	–	–	–	98.75	98.67

Table 2. Recognition rate (%) averaged over all speakers

97.4% lower than at 36dB). Higher accuracy leads to higher sensitivity to variations between training and testing SNR. We observed that this sensitivity is highly noise dependent. For white noises (Gaussian and Hair Dryer), LDA space is very sensitive to SNR variations, especially at low SNRs. For the other noises tested in our experiments, LDA remains efficient to small changes around the training SNR.

ACKNOWLEDGEMENT

The author would like to thank Dr. Yifan Gong and Prof. Jean-Paul Haton, from CRIN–INRIA Lorraine for their help and advice throughout this work.

REFERENCES

[1] R. Haeb-Umbach and H. Ney. Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. In *Proc. IEEE ICASSP*, volume I, pages 13–16, San Francisco, California, March 1992.

[2] R. Haeb-Umbach, D. Geller, and H. Ney. Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities. In *Proc. IEEE ICASSP*, volume II, pages 239–242, Minneapolis, Minnesota, USA, April 1993.

[3] M. J. Hunt and C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. IEEE ICASSP*, pages 262–265. 1989.

[4] M. J. Hunt, D. C. Bateman, S. M. Richardson, and P. Piau. An investigation of PLP and IMELDA acoustic representation and of their potential combination. In *Proc. IEEE ICASSP*, pages 881–884, Toronto, Canada, 1991.

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, 1973.

[6] B. S. Atal. Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, 1972.

[7] Y. Gong and J.-P. Haton. Stochastic Trajectory Modeling For Speech Recognition. In *Proc. IEEE ICASSP*, volume 1, pages 57–60, Adelaide, Australia, April 1994.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from incomplete data via the EM Algorithm. *Journal of Royal Statistical Society Ser. B*, 39:1–39, 1977.

[9] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit, 1992.

[10] S. J. Young. HTK: Hidden Markov Model Toolkit V1.4 reference manual. Technical report, Speech group, Cambridge University Engineering Department, September 1992.