NONLINEAR DISCRIMINANT FEATURE EXTRACTION FOR ROBUST TEXT-INDEPENDENT SPEAKER RECOGNITION

Yochai Konig, Larry Heck, Mitch Weintraub, and Kemal Sonmez

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

RÉSUMÉ

Cet article propose une méthode basée sur l'analyse discriminative non-linéaire pour extraire et sélectionner un ensemble de vecteurs acoustiques utilisés pour l'identification de locuteurs. L'approche consiste à mesurer et grouper un grand nombre de mesures acoustiques (correspondant à plusieurs trames de données consécutives), et à réduire la dimensionalité du vecteur résultant au moyen d'un reseau de neurones artificielles. Le critère utilisé pour optimiser les poids du réseau consiste à maximiser une mesure de la séparation entre les locuteurs d'une base de données d'apprentissage. L'architecture du reseau est telle que l'une de ses couches intermédiaires représente la projection des vecteurs acoustiques d'entrée sur un espace de dimensionalité inferieure. Après la phase d'apprentissage, cette partie du réseau peut etre isolée et utilisée pour projeter les vecteurs acoustiques d'une base de données de test. Les vecteurs acoustiques projetés peuvent alors être classifiés. Combiné à un classificateur cepstral, le classificateur utilisant ces nouveaux vecteurs acoustiques réduit de 15% le taux d'erreur de classification de la base de données définie par NIST en 1997 pour l'évaluation des systèmes de reconnaissance du locuteur.

ABSTRACT

We study a nonlinear discriminant analysis (NLDA) technique that extracts a speaker-discriminant feature set. Our approach is to train a multilayer perceptron (MLP) to maximize the separation between speakers by nonlinearly projecting a large set of acoustic features (e.g., several frames) to a lower-dimensional feature set. The extracted features are optimized to discriminate between speakers and to be robust to mismatched training and testing conditions. We train the MLP on a development set and apply it to the training and testing utterances. Our results show that by combining the NLDA-based system with a state of the art cepstrum-based system we improve the speaker verification performance on the 1997 NIST Speaker Recognition Evaluation set by 15% in average compared with our cepstrum-only system.

1. INTRODUCTION

Our goal is to extract and select features that are more invariant to non-speaker-related conditions such as handset type, sentence content, and channel effects. Such features will be robust to mismatched training and testing conditions of speaker verification systems. With current feature sets (e.g., cepstrum) there is a big performance gap between matched and mismatched tests [8] even after applying standard channel compensation techniques [4]. In order to find these features, the feature extraction step should be directly optimized to increase discrimination between speakers, and to filter out the non-relevant information.

Our proposed solution is to train a multilayer perceptron (MLP) to nonlinearly project a large set of acoustic features to a lower-dimensional feature set, such that it maximizes speaker separation. We train the MLP on a development set that includes several realizations of the same speakers under different conditions. We then apply the learned transformation (MLP in feed-forward mode) to the training and testing utterances. Finally, we use the resulting features for training the speaker recognition system, e.g., Bayesian adapted Gaussian mixture system [9].

We begin by reviewing related studies in Section 2. We describe the proposed feature extraction technique in Section 3. The Development database is described in Section 4. In Section 5, we report the experimental results on the 1997 NIST evaluation set. We continue with analysis of the results in Section 6. Finally, we conclude and describe directions for future work in Section 7.

2. RELATED STUDIES

The related studies to the NLDA technique can be divided into two main categories: robust speaker verification systems, and data-driven feature extraction techniques. Previously proposed approaches to increase robustness to mismatched training and testing conditions, especially to handset variations, include handset-dependent background

models [3], and a handset-dependent score normalization procedure known as Hnorm [9]. Data-driven feature extraction techniques were mainly suggested for speech recognition tasks. Rahim, Bengio and LeCun suggested optimizing a set of parallel class specific (e.g., phones) networks performing feature transformation based on minimum classification (MCE) criterion [7]. Fontaine, Ris and Boite used 2-hidden layer MLP to perform NLDA for isolated word, large vocabulary speech recognition task [2]. The training criterion for the MLPs was phonetic classification. Bengio and his colleagues suggested a global optimization of a neural network-hidden Markov (HMM) hybrid, where the outputs of the neural network constitute the observation sequence for the HMM [1].

3. NONLINEAR DISCRIMINANT ANALYSIS (NLDA)

We explore a nonlinear discriminant analysis (NLDA) technique that finds a nonlinear projection of the original feature space into a lower dimensional space that maximizes speaker recognition performance. This maximization problem can be expressed as

$$A^* = \underset{A}{\text{argmax}} \; J\{A(X)\} \qquad (1)$$

Where $A(X)$ is a nonlinear projection of the original feature space X onto a lower dimensional space, and $J\{\}$ is a closed-set speaker identification performance measure. To find the best A we train a 5 layer multilayer perceptron (MLP) to discriminate between speakers in a carefully selected development set (as described below). The MLP is constructed from a large input layer, a first large nonlinear hidden unit, a small ("bottleneck") second linear hidden layer, a large third nonlinear hidden layer, and a softmax output layer (see Figure 1). The idea is that A is the projection of the input features speaker onto the "bottleneck" layer. After training the 5-layer MLP (denoted 'MLP5') we can remove the last hidden layer and the output layer, and use the remaining 3-layer MLP to project the target speaker data. Then, we use the transformed features to train the speaker verification system, for example, a Bayesian adapted GMM system (see Figure 2). The underlying assumption is that the transformation as found in the development set will be invariant across different speaker populations.

4. DEVELOPMENT DATABASE

To train the 5-layer MLP, we chose 855 Switchboard sentences (about 2 hours) from 31 speakers with a balanced mix of carbon and electret handsets, and balanced across gender. The input consists of 17 cepstral coefficients

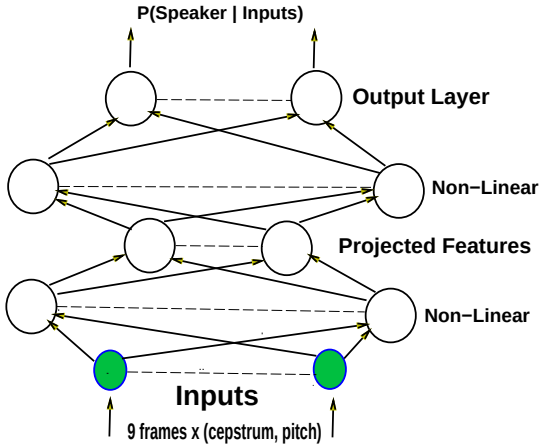

Figure 1: MLP5 for Speaker Recognition

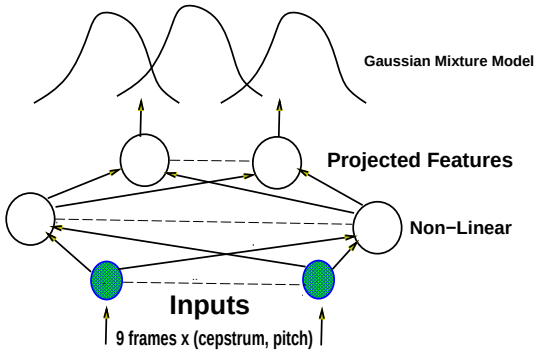

Figure 2: MLP3 for Feature Transformation

and an estimate of the pitch for the current frame, four past frames and four future frames, resulting in a 162-dimension vector. The first hidden layer has 500 sigmoidal units, the bottleneck layer has 34 linear units, the third hidden layer has 500 sigmoidal units, and a softmax output layer has 31 outputs (one for each speaker in the development set). After training the MLP5, we chopped the upper two layers. The resulting MLP ('MLP3') has one hidden layer and was used to transform the data of the target and impostor speakers in a test set as described above.

5. EXPERIMENTAL RESULTS

We used the 1997 NIST Speaker Recognition Evaluation corpus for testing. We report results for three different systems: (1) our best cepstrum system, which is our implementation of the state of the art in text independent speaker verification systems [6]) with 33 input features comprised of 10 cepstral coefficients, energy term, and first and second time derivatives (2) the NLDA based system described in this paper, (3) a combination of the cepstrum and the

Test	Cepstrum	NLDA	Combined
female 3	18.4%	23.0%	16.7%
female 10	12.1%	14.6%	10.8%
female 30	10.5%	12.4%	9.0%
male 3	14.9%	19.4%	14.4%
male 10	13.2%	12.9%	11.1%
male 30	7.9%	11.0%	7.1%

Table 1: Equal Error Rate (EER) Results of the 1997 NIST Eval., 1h condition

Test	Cepstrum	NLDA	Combined
female 10	13.5%	17.0%	12.5%
male 10	11.3%	14.4%	10.5%

Table 2: Equal Error Rate (EER) Results of the 1997 NIST Eval., 1s condition

NLDA systems. The third system is a linear combination of the normalized scores with weights of 0.7 for the cepstrum system scores and 0.3 for the NLDA system scores (expect for the 3 second cases, where we used 0.6 for the cepstrum system and 0.4 for the NLDA system). We use the equal error rate (EER) between misses and false alarms as a performance measure for reporting results. In Table 1, we summarize the results for the 1h condition in the NIST evaluation. In this condition the training consists of 2 phone calls from the same handset, each 1 minute in duration. There are three different test lengths: 3, 10, and 30 seconds. We report the results for each gender separately, by pooling all the test data together (matched and mismatched telephone number).

The results show a consistent win for the combined system over our state of the art cepstrum system. We observe the same consistent win for another condition, 1s, in the 1997 NIST Speaker Recognition Evaluation as demonstrated for the 10 second case in Table 2, and across all regions of the DET (false alarm probability versus miss probability) curves as illustrated in Figure 3 for the male, 10 seconds (1h condition) for the cepstrum only system and the combined system. These results are consistent with our initial results for the 1998 Evaluation corpus.

6. RESULT ANALYSIS

In this section, we examine our “black box” approach, provide insight to its success and give directions for potential improvements. In order to examine the importance of the pitch input, the 9 frame temporal window, and the degradation loss as a result of the dimension reduction

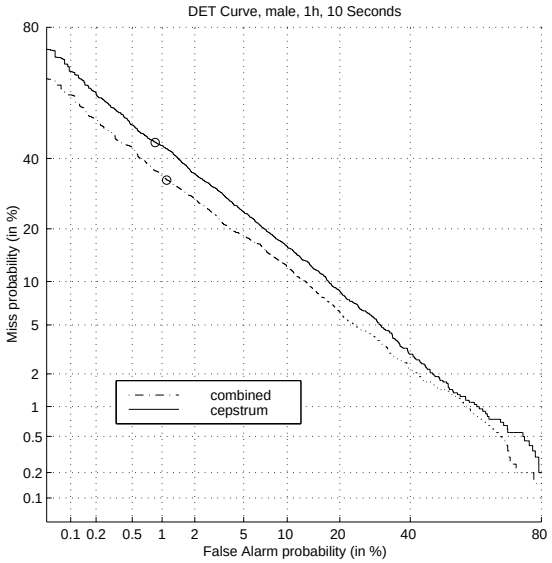

Figure 3: DET Curve for male, 1h, 10 seconds

Inputs	Name	Frame Correct
9 frames + pitch	MLP3	37.2%
9 frames + pitch	MLP5-34	28.9%
9 frames + pitch	MLP5-50	29.0%
9 frames, no pitch	MLP5-NO	25.9%
1 frame + pitch	MLP5-1fr	18.6%

Table 3: Frame-level results on the cross-validation set

from 162 inputs to 34 hidden units in the bottleneck layer, we trained several MLPs and tested their cross-validation, frame-level performance on a close set speaker recognition (our development set as described above). In the development phase we found a strong correlation between these frame-level results and the “full cycle” results of the speaker verification system. The results are summarized in Table 3.

We trained two types of MLPs: a 5-layer MLP, and a "vanilla” MLP with three layers including one hidden layer (denoted ‘MLP3’). As mentioned above there were 31 speakers in our development set, 687156 frames for training and 77904 for cross-validation. Our baseline MLP is the MLP5 described above with 162 inputs and 3 hidden layers with 500, 34, and 500 units (named ‘MLP5-34’). The output layer of all our nets has 31 outputs, one output for each speaker in our development set. The MLP5 named ‘MLP5-NO’ is the same as the baseline but without pitch information (only 153 inputs). The MLP5 named ‘MLP5-1’ is the same as the baseline but with only one input frame (as compared to the 9 frames used in the other systems)

Training a 5-layer MLP is difficult given the complex

nonlinear error surface and requires a lot of training data preferably a ratio of at least 10 between frames than free parameters. In these experiments the ratio was around 4.7 (700k frames to 150k parameters). This might explain the disparity in performance between the MLP3 to the MLP5. This is not due to the bottleneck size as shown by the result of the MLP5 named 'MLP5-50' (the same as 'MLP5-34' but with 50 hidden units in the bottleneck layer). In our speech recognition experiments [5] with NLDA, with the right ratio between frames to free parameters, we did not observe any performance loss because of the dimension reduction at the bottleneck layer. Thus, we plan to increase the size of the development set and hopefully improve the performance of the MLP5 and the overall technique. Additionally comparing the second row to the fourth and fifth rows in Table 3, we observe from these results that that we get a 3% absolute gain from the pitch information, and 10.3% absolute gain from the temporal window.

Another set of interesting results is the correlation between the cepstrum and the NLDA scores on 1997 Eval. set, 1h condition, as summarized in Table 4. From these results, we observe that the NLDA technique contribute a significant amount of new information, especially for the shorter test lengths. This is consistent with the results previously shown in Table 1.

Test Length	Male	Female
3	0.61	0.47
10	0.68	0.71
30	0.76	0.77

Table 4: Correlation Coefficients between NLDA and Cepstrum systems on 1997 Eval. set, 1h condition

7. CONCLUSIONS AND FUTURE WORK

We presented a nonlinear discriminant analysis (NLDA) technique that extracts a speaker-discriminant feature set. Our results on the 1997 NIST evaluation show a consistent (across 12 different tests) and significant (around 15% in relative error) improvement when combining the system trained with the NLDA features with cepstrum based system. Our initial results on 1998 NIST evaluation are consistent with 1997 results. Furthermore, our analysis suggests that there is a potential for performance improvement given more development data. We also plan to experiment with other types of input data such as speech over cellular phones and speaker-phone speech. In addition, we plan to extend this study by using a wider range of input representations and resolutions such as first and second derivatives of cepstrum, filter-bank energy levels, and different analysis windows. Finally we want to note that although the training of the MLP with 5 layers is computationally expensive (25 x real time), the application of the MLP3 in a feed forward mode is very fast (less than 0.4 real-time), thus the NLDA approach is feasible in realistic settings.

8. REFERENCES

[1] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global optimization of a neural network-hidden Markov model hybrid. *IEEE trans. on Neural Networks*, 3(2):252–258, March 1992.

[2] V. Fontaine, C. Ris, and J. M. Boite. Nonlinear discriminant analysis for improved speech recognition. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Rhodes, Greece, 1997.

[3] L. P. Heck and M. Weintraub. Handset-dependent background models for robust text-independent speaker recognition. In *Proceedings International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Munich, Germany, 1997.

[4] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, pages 1367–1370, 1991.

[5] Y. Konig and M. Weintraub. Acoustic modeling session - SRI site presentation. In *NIST LVCSR Workshop*, Linthicum Heights, MD, October 1996.

[6] NIST. Result summary. In *Speaker Recognition Workshop Notes*, Linthicum Heights, Maryland, 1997.

[7] M. Rahim, Y. Bengio, and Y. LeCun. Discriminative feature and model design for automatic speech recognition. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Rhodes, Greece, 1997.

[8] D. A. Reynolds. The effects of handset variability on speaker recognition performance experiments on the switchboard corpus. In *Proceedings International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Atlanta, GA, 1996.

[9] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Rhodes, Greece, 1997.