論文 / 著書情報
Article / Book Information

Title	A Weight Estimation Method Using LDA for Multi-Band Speech Recognition
Authors	Koji Iwano, Kaname Kojima, Sadaoki Furui
Citation	Interspeech 2006, Vol. , No. , pp. 2534-2537,
Pub. date	2006, 9
Copyright	(c) 2006 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

A Weight Estimation Method Using LDA for Multi-Band Speech Recognition

Koji Iwano, Kaname Kojima, and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1-W8-77 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{iwano, kaname, furui}@furui.cs.titech.ac.jp

Abstract

This paper proposes a band-weight estimation method using Linear Discriminant Analysis (LDA) for multi-band automatic speech recognition (ASR). In our scheme, a spectral domain feature, SPEC, is modeled using a multi-stream HMM technique. This paper also proposes the use of Output Likelihood Normalization (OLN) in combination with the LDA-based weight-estimation method in order to adjust the relative weights of individual word (phoneme) models. Experiments were conducted using Japanese connected digit speech in various kinds of noise and SNR conditions. Experimental results show that the proposed LDA-based method is effective in all noise conditions. The results also confirm that the combination of OLN with the LDA-based method further increases noise robustness of the multi-band ASR. Furthermore, comparing the results of LDA applied to the SPEC and MFCC features respectively, it can be seen that greater performance gains are achieved with the former case than with the latter; this means that SPEC within a multi-band speech recognition framework can more effectively deal with the noise contamination than MFCC.

Index Terms: multi-band speech recognition, band-weight estimation, linear discriminant analysis (LDA), spectral domain feature.

1. Introduction

In most state-of-the-art speech recognition systems, speech is converted into a time-signal of the MFCC vector. However, this method tends to spread any noise the system encounters across all the MFCC coefficients, even when the contaminating noise is confined to a narrow frequency band. This shortcoming makes it substantially more difficult to develop effective methods to eliminate the effects of noise contamination. The application of frequency-band dependent processes to a multi-band ASR approach is expected to be more effective for developing noise-robust ASR systems [1, 2, 3, 4, 5].

For multi-band ASR, accurately estimating sub-band weights according to the reliability of individual spectral bands is one of the key issues. In previous work, estimated signal-to-noise ratio (SNR) for each band [1, 2, 3], entropy for each band [3], an output of multi-layer perceptron [4], or the maximum likelihood (ML) criterion [6] was used for determining the sub-band weights. In this paper we propose the use of Linear Discriminant Analysis (LDA) as a new and effective means for estimating the reliability weights and evaluate the feasibility of this approach in the framework of multi-band ASR in various noise environments. We also apply Output Likelihood Normalization (OLN) [7], which was originally proposed as a weight-normalization method for audio-visual speech recognition, to our system and evaluate its effects when combined with the LDA-based weight-estimation method.

This paper is organized as follows. Section 2 explains the spectral domain feature used in our multi-band recognition system. In Section 3, our multi-band recognition scheme using the multi-stream HMM technique is explained. Section 4 proposes a weight-estimation method using the LDA and combination of the LDA-based weight-estimation method with the OLN method. Experimental results are presented in Section 5, and Section 6 concludes this paper.

2. Spectral domain features (SPEC)

For the purpose of conducting multi-band speech recognition, a spectral domain feature is used, which will be referred to as "SPEC" hereafter. Although, unlike the MFCC, the SPEC is normalized only in the spectral domain, both features are theoretically similar. Figures 1 and 2 show the feature extraction flows of MFCC and SPEC, respectively. Below is a series of simple descriptions, which illustrates the differences between the two features.

2.1. Mean log-energy subtraction

A spectral bias component (C_0 in the MFCC domain) is removed. Since the absolute value of the energy is changeable depending on many factors including recording conditions, it needs to be normalized.

2.2. Spectrum peak emphasis

Since spectral peaks convey important information for speech recognition, the liftering process for MFCC is effective for raising the recognition performance. For extracting SPEC, the log-spectrum is passed through an FIR filter, with the formula described below, in order to emphasize its spectral peaks and valleys.

$$H(z) = 1 - pz^{-1} \tag{1}$$

2.3. Log-spectral mean subtraction

The time-average of the log-spectrum is subtracted from the log-spectrum of each frame. This process corresponds to Cepstrum Mean Subtraction (CMS) in the cepstral domain. Like CMS, this process is effective for normalizing spectral variations due to transmission characteristics and voice individuality.

3. Multi-band speech recognition

Multi-band ASR is a technique which uses multiple frequency bands for recognition. Multi-stream HMMs provide a means of accomplishing weighting for each band. Although otherwise similar

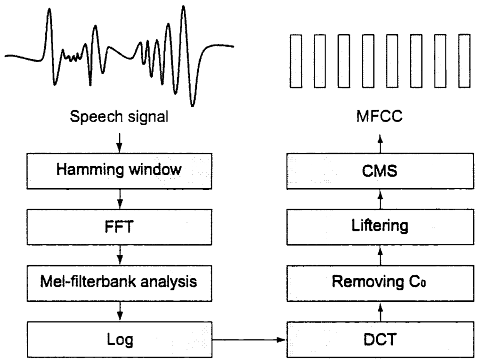

Figure 1: Flow for the MFCC extraction process.

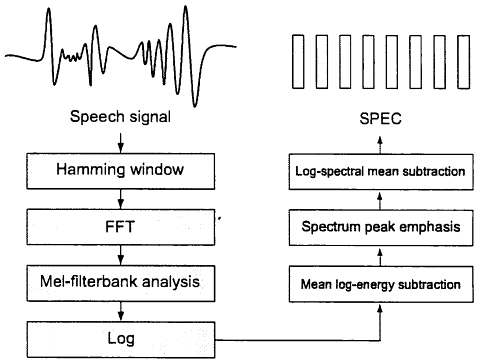

Figure 2: Flow for the SPEC extraction process.

to a standard HMM, log-likelihoods for multi-stream HMMs are calculated in a slightly different manner. When a t-th frame value o_t is observed, the log-likelihood for the multi-stream HMM, $b(o_t)$, is calculated as the weighted sum of the likelihood of each stream, $b(o_{st})$, as indicated by the formula below:

$$b(o_t) = \sum_{s=1}^{S} \lambda_s \cdot b(o_{st}) \tag{2}$$

where S is the number of streams and λ_s represents the weight of an individual stream s.

In order to weight individual bands of SPEC, each dimension of SPEC is considered as a separate stream and thus fed as an input to the multi-stream HMM. This is made possible by the paradigm of the multi-stream HMM. Although a feature vector for recognition actually includes ΔSPEC and Δ log power components, these Δ terms are treated as a single vector stream.

4. Likelihood weight estimation

Linear Discriminant Analysis (LDA) is used to estimate the relative weights of individual streams. In addition, Output Likelihood Normalization (OLN) is used in order to adjust the relative weights for individual models. The weight for the stream containing the Δ terms is fixed to 1.0.

4.1. Stream weight estimation using Linear Discriminant Analysis (LDA)

As discussed above, the log-likelihood value is calculated as the linear sum of the weighted log-likelihood of all individual streams. Since a discriminant function obtained using the LDA has the form of a linear sum, the LDA can be directly used to estimate the stream weights. This maximizes the system's ability to discriminate correct and incorrect input words, and thereby estimates the reliability of each individual frequency band.

First forced alignment is carried out with the training data. Word (phoneme) strings $w_1, w_2, \ldots, w_N$ (N denotes the number of phoneme strings) and their corresponding labels are prepared from the training data. Word (phoneme) labeling is carried out with an automatic labeling process before likelihood estimation. Next, the feature vector o_{w_n}, corresponding to word (phoneme) w_n, is fed as input to all unweighted models m_v ($v = 1, 2, \ldots, V$, where V represents the total number of different models). For each stream s, frame-averaged log-likelihoods are plotted as the coordinate x_s (in an s dimensional space). After plotting all w_n-m_v combinations, LDA is used to obtain a linear discriminant function which is then applied to separate the correct ($w_n = m_v$) from the incorrect ($w_n \neq m_v$) distributions. The obtained linear discriminant function is:

$$a_0 + \sum_{s=1}^{S} a_s x_s = 0. \tag{3}$$

When the LDA produces negative coefficients, which are undesirable for our purpose, they are converted to 0.

$$a_i' = \begin{cases} a_i & (a_i \geq 0) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Finally, the mean value of the weighting coefficient, λ_s, is normalized using the following equation:

$$\lambda_s = S \cdot \frac{a_i'}{\sum_j a_j'}. \tag{5}$$

4.2. Word model weight adjustment using Output Likelihood Normalization (OLN)

Because these stated likelihoods tend to vary widely in low SNR environments, and since this can be a major factor in bringing down recognition quality, reducing this variation affords the possibility of a considerable improvement in performance. In order to help minimize the variation, in OLN, estimation data is used to calculate the average likelihood of unit segments which correspond to individual models, and the inverse of this average is then used to weight the given model [7]:

$$\bar{b}_{m_v} = \sum_{n=1}^{N} b_{m_v}(o_{w_n})/T_A \tag{6}$$

where $b_{m_v}(o_{w_n})$ represents the log-likelihood of a given feature vector o_{w_n}, corresponding to word (phoneme) w_n, and word (phonene) model m_v. $\bar{b}_{m_v}$ thus represents the average log-likelihood for model m_v, N represents the number of segmented features, and T_A represents the total frames for said features. The inverse of $\bar{b}_{m_v}$ is used as the model weight, after normalizing the mean weight to 1.0. The final model weight is thus calculated as follows:

$$\lambda_{m_v} = V \cdot \frac{1/\bar{b}_{m_v}}{\sum_{v=1}^{V} 1/\bar{b}_{m_v}} \tag{7}$$

where V represents the number of models. In order to incorporate the case where both stream weights and model weights are considered simultaneously, a combined, two-dimensional weight λ_{s,m_v} comprising both s and m_v is constructed:

$$\lambda_{s,m_v} = \lambda_s \cdot \lambda_{m_v}. \tag{8}$$

5. Experiments

5.1. Speech data

The speech data used for training and recognition consisted of Japanese continuous digit utterances from 11 male speakers, recorded in a clean environment. Each speaker recorded a series of 210 digit strings, each consisting of 2-8 digits, resulting in a total 1050 digit-utterances per speaker. The "leave one out" method was used; that is, the clean recorded utterances of 10 speakers were used for training, and the utterances from the remaining speaker were used for testing. This process was carried out 11 times, and the recognition rates for the 11 experiments were averaged.

The elevator hall noise, train station noise, and in-car noise from the noise database distributed by the Japan Electronic Industry Development Association (JEIDA) [8] were used as the source of noise contamination for the trials. The former two noises can be classified as "babble noises". The noises were numerically added to testing data at respective SNRs of 5, 10, and 20dB. All waveforms were sampled at 16kHz with 16bit resolution.

In order to investigate baseline performance of the proposed method, data for the weight estimation were created by adding noise signals at the same conditions as testing to the clean training data. Phoneme labeling information used for the estimation was obtained via Viterbi alignment using correct phoneme sequences and the original clean data.

5.2. Acoustic feature vectors

The SPEC-based acoustic feature vector consisted of 27 dimensions, comprising 13 SPEC, 13 Δ SPEC and one Δ log power. Since the number of frequency bands was set at 13, the normalization processes were applied to the 13-dimensional spectral vectors. Acoustic analysis was conducted with 25ms frame windows and a frame shift of 10ms. Log-spectral mean subtraction was applied to each speech file containing 10 connected digit utterances by a single speaker.

5.3. Modeling

Three-state triphone HMMs were first trained as single stream HMMs by using HTK [9]; that is, all streams were jointly trained. The observation probability density for each state was represented by 4-mixture Gaussian distributions. A diagonal covariance matrix was used for each distribution.

Then, all feature vectors in the HMMs were separated into 14 vector streams as shown in Figure 3. All the initial stream weights were set to 1.0.

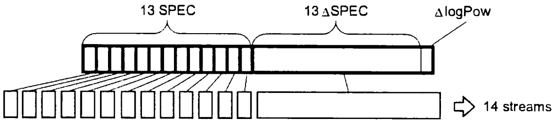

Figure 3: Separation of a feature vector for building multi-stream HMMs. A feature vector of 13 SPEC was separated into 13 streams, and a vector consisting of 13 Δ SPEC and Δ log power was treated as a single stream.

5.4. Experimental results

Four different likelihood weight estimation methods were evaluated. In the first experiment (NONE), the original model, with all likelihood weights initialized to 1.0, was used; no estimation or adjustment method was applied in this case. In the second experiment (LDA), weights were estimated exclusively with Linear Discriminant Analysis; in the third experiment (OLN), weight adjustment between triphone HMMs was conducted exclusively with Output Likelihood Normalization; in the fourth experiment (LDA+OLN), LDA was used to obtain preliminary weight estimations, and weight adjustment across HMMs was then carried out with OLN. In these experiments, both LDA and OLN were applied at the word level.

Since the recognition task targeted continuous digit utterances, a network grammar, consisting of every combination of digits 0-9, was used for language modeling. The most salient insertion penalty was experimentally chosen for each experiment.

The digit recognition results for these four methods, as applied to the SPEC approach, can be seen in Table 1. As can be seen from the results, LDA proves an effective means of estimating likelihood weights; the digit accuracies are highly improved by LDA from the baseline results (NONE) in all noise conditions. The number of spectral bands, for which negative coefficients were produced by LDA and converted to 0 using Eq. (4), varied according to noise conditions from 0 (clean and in-car noise at 20dB) to 2.2 (station noise at 5dB) in average. The effectiveness is also obtained when applying OLN-based weight adjustment. The results also confirm that both LDA and OLN methods applied in concert are more effective than either one alone.

As a supplementary experiment, we compared the relative robustness of the SPEC-based approach with the MFCC-based approach for confirming the effectiveness of spectral-domain features. The MFCC-based acoustic vector consisted of 25 dimensions, and comprised 12 MFCC, 12 Δ MFCC, and one Δ log power. Both the SPEC and MFCC-based acoustic feature vectors have the same freedom of 25 dimensions after normalization. In the MFCC case, the corresponding multi-stream system, like the SPEC-based recognition scheme, Δ terms were treated as a single vector stream and not used for weight estimation. Figure 4 shows the digit accuracies for MFCC-based multi-stream speech recognition and SPEC-based multi-band speech recognition when applying the LDA-based weight-estimation method. Results by MFCC-based recognition without any weighting are also shown in the figure. By comparing these results with those obtained by weighted MFCC, it can be seen that the LDA-based weighting is also useful for MFCC. However, by comparing the results by weighted MFCC and weighted SPEC, it can be seen that greater performance gains are achieved with the SPEC-based features than with the MFCC-

Table 1: Comparison of digit accuracies (%) for four kinds of weight-estimation methods (NONE, LDA, OLN, and LDA+OLN) in various noise conditions. These methods were applied to the SPEC-based multi-band speech recognition.

method	clean	elevator hall noise			train station noise			in-car noise		
		20dB	10dB	5dB	20dB	10dB	5dB	20dB	10dB	5dB
NONE	99.6	91.4	51.0	30.5	88.4	36.1	23.2	98.9	83.0	64.3
LDA	99.3	94.6	59.9	39.3	92.7	52.7	33.7	99.2	94.9	84.2
OLN	99.6	93.1	61.1	39.7	91.0	51.3	32.6	99.1	95.3	84.9
LDA+OLN	99.2	94.7	64.8	41.4	93.3	55.8	35.6	99.0	95.4	87.4

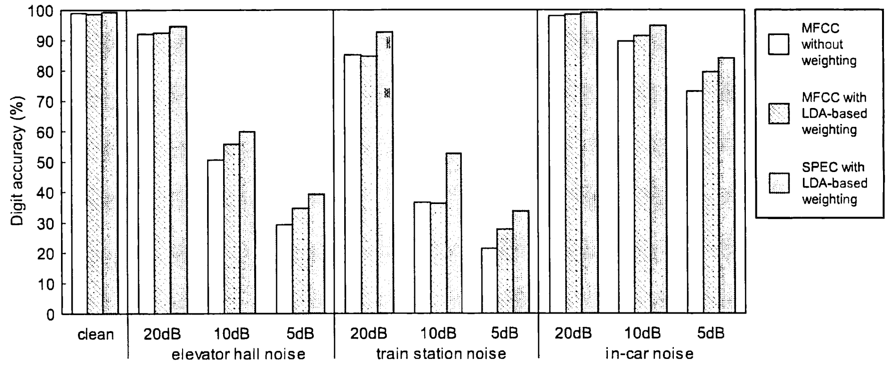

Figure 4: Comparison of digit accuracies by MFCC without weighting, MFCC with LDA-based weighting, and SPEC with LDA-based weighting.

based features in all noise conditions. This fact bears out our hypothesis with respect to the utility of controlling the spread of noise contamination across frequency bands.

6. Conclusions

This paper proposed the use of Linear Discriminant Analysis (LDA) for obtaining favorable likelihood weights in the context of spectral-based multi-band speech recognition. Experimental results using connected digit speech recognition show that the proposed LDA-based weight-estimation method is effective in various noise conditions. It was also confirmed that the combination of Output Likelihood Normalization (OLN) with the LDA-based method further increases noise robustness. We further showed that within the context of a multi-band approach, the proposed method affords greater gains in noise robustness to spectral (SPEC), rather than cepstral (MFCC) based approaches.

The current experiment was conducted with explicit knowledge about the SNR values, and phoneme labeling for the estimation data, and this being a success, the next step is to prove the feasibility of our approach in a real-world setting, where these variables would not be explicitly available to the system. It is necessary to confirm the utility of incorporating spectral subtraction, noise adaptation techniques such as MLLR, or the multi-condition training into the current approach. Finally, it is also necessary to compare the performance of the proposed method with other multi-band ASR approaches.

7. References

[1] Bourlard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP96*, Philadelphia, PA, USA, vol.1, pp.426–429, October 1996.

[2] Hermansky, H., Tibrewala, S., and Pavel, M., "Towards ASR on partially corrupted speech," *Proc. ICSLP96*, Philadelphia, PA, USA, vol.1, pp.462–465, October 1996.

[3] Okawa, S., Bocchieri, E., and Potamianos, A., "Multi-band speech recognition in noisy speech," *Proc. ICASSP98*, Seattle, WA, USA, vol.2, pp.641–644, May 1998.

[4] Morris, A., Hagen, A., Glotin, H., and Bourlard, H., "Multi-stream adaptive evidence combination for noise robust ASR," Speech Communication, vol.34, nos.1-2, pp.25–40, April 2001.

[5] Ming, J. and Smith, F.J., "Union: A new approach for combining sub-band observations for noisy speech recognition," Speech Communication, vol.34, nos.1-2, pp.41–55, April 2001.

[6] Hagen, A., Bourlard, H., and Morris, A., "Adaptive ML-weighting in multi-band recombination of gaussian mixture ASR," *Proc. ICASSP2001*, Salt Lake City, UT, USA, vol.1, pp.257–260, May 2001.

[7] Tamura, S., Iwano, K., and Furui, S., "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," *Proc. ICASSP2005*, Philadelphia, PA, USA, vol.1, pp.469–472, March 2005.

[8] http://www.sunrisemusic.co.jp/dataBase/fl/noisedata01_fl.html

[9] http://htk.eng.cam.ac.uk/