MINIMUM KULLBACK-LEIBLER DISTANCE BASED MULTIVARIATE GAUSSIAN FEATURE ADAPTATION FOR DISTANT-TALKING SPEECH RECOGNITION

Yue Pan and Alex Waibel

Interactive System Laboratories, Carnegie Mellon University, Pittsburgh, PA, USA
Email: {ypan, waibel}@cs.cmu.edu

ABSTRACT

Multivariate Gaussian based speech compensation or mapping has been developed to reduce the mismatch between training and deployment conditions for robust speech recognition. The acoustic mapping procedure can be formulated as a feature space adaptation where input noisy signal is transformed by a multivariate Gaussian network. We propose a novel algorithm to update the network parameters based on minimizing the Kullback-Leibler distance between the core recognizer's acoustic model and transformed features. It is designed to achieve optimal overall system performance rather than MMSE on a specific feature domain. An online stochastic gradient descent learning rule is derived. We evaluate the performance of the new algorithm using JRTk Broadcast news system on a distance-talking speech corpus and compare its performance with that of previous MMSE based approaches. The experiments show the KL based approach is more effective for a large vocabulary continuous speech recognition (LVCSR) system.

1. INTRODUCTION

Automatic speech recognition systems attain high performance for close-talking applications. However, they degrade significantly in adverse distant-talking environment. The reason is the mismatch between training and deploying condition caused by room reverberation and environmental noise. Two major classes of approaches, model domain and feature-domain adaptation have been studied to address the problem. Model domain adaptation (e.g. MLLR [1]) directly modifies the recognizer's HMM parameters to match statistics of the received sensor signal. It is computationally costly for a LVCSR system (more than 100,000 Gaussians in our setup). Feature domain adaptation aims at compensating the speech feature distortion due to the interfering noise. It is simpler and computational affordable.

In this paper, we introduce a novel optimized feature adaptation measure to increase the robustness of an LVCSR system. Our approach assumes a mixture of multivariate Gaussian (GMM) acoustic model to represent received speech. A family of multivariate Gaussian based feature compensation techniques have been developed, such as RATZ [2], and SPLICE[3]. They are designed to minimize the mean square error (MMSE) between processed ASR feature computed from distorted speech and clean speech. The new feature adaptation algorithm KLFA that we propose is based on minimum Kullbak-Leibler distance instead. One main advantage is the optimization object function is defined based on the core recognizer's HMM output probability as opposed to relying on the less precise general acoustic model. Therefore, the learning process takes full advantage of the feedback from the system output, and leads to optimal overall performance. Although our multivariate Gaussian network is applied on a frame-by-frame basis, it can take into account the correlation effects of the neighbor feature, since the parameter learning process is affected by the core recognizer's time delay linear transform. The effectiveness of the proposed algorithm is demonstrated in our LVCSR experiments, and it is shown to be consistently superior to previous MMSE based approach.

In the remaining of this paper, we will describe the procedure of multivariate Gaussian based feature adaptation; then, the proposed minimum KL distance estimation based KLFA algorithm is presented; and finally, KLFA and previous techniques will be evaluated and compared using LVCSR system with additive noisy speech and real distant-talking speech.

2. MULTIVARIATE-GAUSSIAN-BASED COMPENSATION

In distant-talking environment, clean speech spectral coefficient x is filtered by a room channel h and corrupted by additive stationary background noise n:

$$y_{lin} = x_{lin} \cdot h_{lin} + n_{lin} \quad (2.1)$$

where y is the corresponding noisy speech spectral coefficient. In log-spectral domain, the relation becomes non-linear and can be expressed as:

$$y = x + h + \log(1 + e^{n-x-h}) \tag{2.2}$$

We assume the probability density function (PDF) of log-spectra clean speech *p(x)* to be a mixture of multivariate Gaussians:

$$p(x) = \sum_{k=1}^{K} p[k] N(\mu_k, C_k) \tag{2.3}$$

where *p[k]*, μ_k and C_k represent respectively the *a prior* probabilities, mean vector and covariance matrix of each multivariate Gaussian element *k*. These parameters are determined by using traditional maximum likelihood methods from the training dataset.

We further assume the PDF of the corresponding log-spectral noisy speech *p(y)* is a mixture of multivariate Gaussians as well and the effects of the environment on the statistics of clean speech can be approximated as additive compensations for mean vectors and covariance matrices. Various methods can be used to estimate the parameter of *p(y)* given *p(x)*, such as, vector Taylor series (VTS) and parallel model combination (PMC). When stereo recordings for both the clean and noisy speech data are available as in RATZ, *p(y)* can be directly estimated by traditional EM algorithm from the training dataset.

Knowing the noisy speech model *p(y)*, for one given mixture component *k*, the estimate of clean speech *x* has the form

$$\hat{x}_k = \alpha \cdot y + r_k \tag{2.4}$$

α is an uncertainty weight, and usually set to 1; r_k is the correction vector, and can be learned from noisy speech model. In the case of stereo recording is available, the MMSE estimation of r_k is computed as:

$$r_k = \frac{\sum_t p(k \mid y_t)(x_t - y_t)}{\sum_t p(k \mid y_t)} \tag{2.5}$$

where the summation is over all frames of the stereo training data.

Once the parameters of the distribution of noise speech and the corrections vector codebook are obtained, MMSE estimator is used again to calculate the clean speech given the observed noisy speech vector *y*.

$$\hat{x} = \int x p(x \mid y) dx \approx y + \sum_{k=1}^{K} p(k \mid y) \cdot r_k \tag{2.6}$$

The posterior probability is computed using the mixture of Gaussian model for input noise speech feature.

In summary, the algorithm works in three stages: estimation of the statistics of noisy speech, estimation of the correction vector codebook and compensation of noisy speech. The acoustic feature compensation process can be formulated as a one-layer network consisted of Gaussian nodes in the GMM model and a correction vector codebook. The parameters of this network are trained to maximize the GMM output probability and minimize the mean square error between the clean speech and compensated noisy speech. The transformation performed is a shift by the sum of all correction vector weighted by *P(k|y)*.

Although it is possible to directly take recognizer's speech acoustic model as the general acoustic model, in practice, for a LVCSR system, we have to train an independent smaller GMM model to avoid the complicated ASR feature and decrease the amount of parameters. Using a secondary acoustic model requires much less adaptation data.

3. MINIMUM KULLLBACK-LEIBLER DISTANCE ESTIMATION

Generalizing the multivariate Gaussian network as a transforming function *g(x)*, we pass a *n*-dimension input random variable *x* through *g(x)* to give a n-dimension output variable *y*. The PDF of *y*, $p_y(y)$, can be written as a function of the PDF of the input, $p_x(x)$,

$$p_y(y) = \frac{p_x(x)}{|J|} \tag{3.1}$$

where $|J|$ is the absolute value of the Jacobian of the transformation. The Jacobian is the determinant of the matrix of partial derivatives:

$$J = \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} \tag{3.2}$$

In the case of multivariate Gaussian network, the transform parameters include mean vectors μ_k, covariance matrices C_k, correction vector r_k, and uncertainty weight α, denoted collectively as W:

$$W = \{\mu_1, \cdots, \mu_K, C_1, \cdots, C_K, r_1, \cdots, r_K, \alpha\} \tag{3.3}$$

An error function must be defined first in order to estimate the transform parameters. Instead of using MMSE criterion, we choose the Kullback-Leibler (KL) distance function, which measures the difference between core recognizer's HMM model λ and transformed feature *y*, defined by

$$\varepsilon(W) = \int p_y(y) \log \frac{p_y(y)}{p(y \mid \lambda)} dy = -E[\log p(y \mid \lambda)] - H(y) \tag{3.4}$$

Where the operator *E* performs averaging over the processed feature *y*. As is well known, the KL distance is always non-negative, and is zero only when $p(y|\lambda)=p(y)$.

The error consists of two terms: the first is the negative log-likelihood of transformed feature *y* given the HMM model λ. The second term is the entropy of *y*. *H(y)* is given by:

$$H(y) = -\int p_y(y) \log p_y(y) dy \tag{3.5}$$

substituting (3.1) into (3.5) gives

$$H(y) = E[\log \mid J \mid] - E[\log \; p_x(x)] \quad (3.6)$$

The second term on the right (the entropy of observed feature of sensor signals x) may be considered to be unaffected by alterations of parameter W determining $g(x)$.

Therefore in order to minimize the error by changing W, we need only concentrate on two terms, $E[\log p(y|\lambda)]$ the average log probability of output feature, and, $E[\log|J|]$, the average log of how the input affects the output. Considering x's the adaptation set to approximate the density $p_x(x)$, this can be done by and deriving an online, stochastic gradient descent learning rule:

$$\Delta W \propto \frac{\partial \varepsilon(W)}{\partial W} = \frac{\partial}{\partial W}(-\log p(y \mid \lambda) - \log \mid J \mid) \quad (3.7)$$

During the learning phrase, a convergence algorithm is used on training data to update the parameters in a manner to minimize the error function.

1. Initialize the multivariate Gaussian network's parameters using MMSE based method.
2. Assign a core recognizer's HMM state to each training frame (labeling) by carrying out HMM Viterbi forced alignment algorithm on given reference texts. The first term of the error function will be calculated according to the labels.
3. For each iteration n, forward propagate training data through the transformation network and core recognizer to compute value of the error function; evaluate the partial derivatives for all parameters using back propagation; update the transform parameters by a learning rate η as $W(n+1) = W(n) - \eta \cdot \frac{\partial \varepsilon(W)}{\partial W}$.
4. Repeat step 3, until converging or completing a specific iteration number.

The minimum KL distance estimation has interesting relations to other techniques. Maximum likelihood linear regression (MMLR) based feature space adaptation can be viewed as a special case of KLFA, where the $g(x)$ is a global linear transformation matrix A, with the constraint $|J|=|A|=1$. So only the first term in (3.7) matters, which reduces the problem to maximum likelihood estimation. Another interesting case is its relation to information maximisation based independent component analysis, which does not assume any knowledge of probability distributions and the goal is to maximize entropy of y, i.e., only the second term in (3.7) is considered.

For KLFA with a multivariate Gaussian network topology, $g(x)$ is a non-linear transformation defined on the framed based Mel-scale log-spectrum domain. State-of-the-art LVCSR system employs temporal processing by combining adjacent frames of feature vectors and multiplying a Linear Discriminant Analysis (LDA) transform matrix. Its process can be viewed as a Time Delay Network (TDN), which filter the time trajectories. During the descent learning, the temporal processing is automatically incorporated through the LDA matrix and affects the learning of parameters W. Thus the network takes into account of adjacent frames. The feature adaptation network topology is shown in Figure. 1.

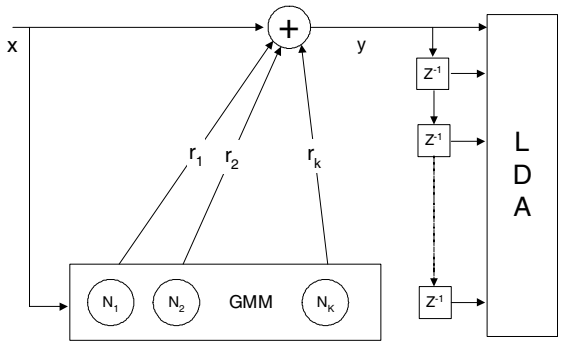


Figure 1. Feature adaptation networks topology

4. SPEECH RECOGNITION EXPERIMENTS

LVCSR experiments are carried out to evaluate KLFA algorithm using both simulated additive noisy speech and real distant-talking speech. The results are compared with MMSE based multivariate Gaussian compensation RATZ, MLLR based feature space adaptation MLLR-FA and a retrained system built on mixing data of clean speech and distant-talking speech.

4.1. Dataset and system

4.1.1. Simulated noisy and real distant-talking speech

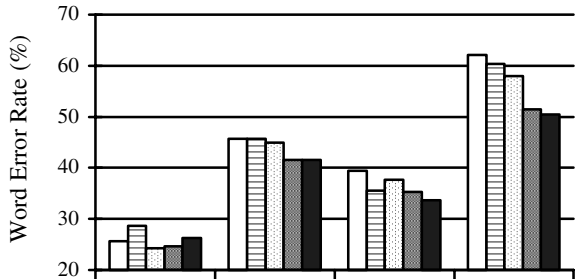
A distant-talking speech corpus contains data collected in ISL Lab using a multi-channel sound card. The 16-bit audio signal is sampled at 16kHz. During recording, the speaker wears a close microphone for reference clean speech signal. Two distant microphones are placed at about 6 feet away: one is directional and pointing towards the speaker, so that it can focus on the signal from the speaker's direction; the other one is un-directional, so that it can detect the sound sources from all direction. Artificial additive noisy speech is also simulated by adding the recorded background noise to close-talking speech at different SNR, specifically, 10dB and 5dB. Speakers read articles chosen at random from a collection of News-hour database. A training set has 3-hour speech in total including both directional and unidirectional distant data. It could be used for retraining baseline system. A test set includes over 600 words articles by one male speaker and one female speaker. An adaptation set

includes 3 minutes speech of the same speakers with given reference texts.

4.1.2. JRTk Broadcast News baseline system

Our HMM LVCSR baseline system is trained on Broadcast News (BN) corpus using JRTk [4]. The core recognizer deploys a tri-phone acoustic model and has over 100K Gaussians. Speech signal is windowed with a 16-ms Hamming window every 10 ms. After Vocal Tract Length Normalization (VTLN), log-spectrum is calculated and averaged into 30 triangular bins arranged at equal Mel frequency intervals. A follow-up LDA matrix is applied on a 7-frame context window and reduces feature dimensionality to 42. The advantage of using LDA transform instead of DCT transform is demonstrated in [4]. Mean Normalization is applied at the final step. The system has a vocabulary size of 40K and a tri-gram language model trained using the Broadcast News corpus. Word error rate of the baseline system is 17.6% on close-talking speech.

4.2. Experimental Results



	SNR 10dB additive noise	SNR 5db additive noise	directional distant mic.	undirection-al distant mic.
Baseline	25.6	45.7	39.4	62.1
MLLR-FA	28.7	45.7	35.6	60.4
RATZ	24.2	45.0	37.7	58.0
KLFA	24.6	41.5	35.3	51.5
Retraining	26.3	41.5	33.6	50.5

Table 1. WER for different noise conditions using various types of algorithms

Various approaches are implemented for comparison. RATZ is an MMSE based multivariate Gaussian adaptation. Its general GMM acoustic model and correction vector codebook are trained using the stereo recording in the adaptation set. Supervised MLLR based feature space adaptation (MLLR-FA) is a popular technique. A 42-dimenstion LDA feature global linear transformation matrix is derived on the adaptation set without using stereo recording. The accurate HMM state path is calculated according to reference texts by Viterbi forced alignment. KLAM is performed under the same conditions as in MLLR-FA. It has the exactly same mapping network structure as RATZ. The GMM network has 100 Gaussians with diagonal covariance matrix and 100 correction vectors. The input and output are 30-dimension Mel-scale log-spectral coefficients. Some simplification is used to speed up KLFA computation. Only the Gaussian component with the highest probability is considered during the mapping process, and then the computation of $|J|$ could be simplified by ignoring the off-diagonal elements. The retrained system is built by using data from both clean speech training set and 3 hour distant-talking speech training set. Its WER for the close-talking speech increases to 22.8%.

The experiment results for these approaches under different condition are summarized in Table 1. It can be seen that KLFA algorithm performs better than all of the previous adaptation algorithms and is comparable to retrained system of mixing data.

5. DISCUSSIONS AND CONCLUSION

Experiments show that the proposed KL distance based error function is more effective for LVCSR system and outperforms the previous approaches. KLFA has the several advantages for a LCVSR system. First, although it use a secondary acoustic model, the estimation criterion is based on feedback from the core recognizer, which is a more reliable to yield the best overall system performance than the MMSE based optimization. Second, it only requires an accurate HMM state path alignment of reference texts, and thus can work without the use of stereo data. Third, with the time-delay LDA transform, it takes account correlation effects of neighbor frames. Last, the nonlinear multivariate Gaussian adapting network is more powerful than a simple linear transform matrix as in MMLR-FA.

6. REFERENCES

[1] C.J. Legetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, Academic Press, Vol. 9, pp. 171-185, 1995.

[2] P.J. Moreno, B. Raj, E. Gouvea and R.M. Stern, "Multivariate-Gaussian-Based Normalization for Robust Speech Recognition", Proc. *ICASSP*, 1995.

[3] L. Deng, A. Acero, M. Plumpe, and X.D. Huang, "Large-vocabulary environment estimation for FCDCN in a continuous speech recognition system," Proc. *ICSLP*, 2000.

[4] H. Yu, and A. Waibel, "Streamlining the Front End of a Speech Recongizer", Proc. *ICSLP*, 2000.