On the Interaction Between Speaker Normalization, Environment Compensation, and Discriminant Feature Space Transformations

*Richard Rose, Alireza Keyvani, and Antonio Miguel**

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada
(rose,keyvani)@ece.mcgill.ca

* Communication Technologies Group (GTC)
I3A, University of Zaragoza, Spain
amiguel@unizar.es

Abstract

This paper presents a study of the interaction between frequency warping based speaker normalization algorithms, environment compensation algorithms, and discriminant feature space transformations (DFT) in providing consistent reductions in ASR word error rate (WER) over a range of acoustic degradations. Performance improvements obtained using speaker normalization algorithms, including vocal tract length normalization (VTLN) and a newly proposed augmented state space acoustic decoder, are shown to improve substantially when applied in a discriminant feature space where acoustic environment compensation has been applied. Furthermore, the effects on ASR performance of the DFT are also shown to be enhanced by reducing within class variability by applying the DFT on a speaker and an environment normalized feature space.

1. Introduction

Many feature space normalization and acoustic model adaptation techniques that are specifically designed to compensate for speaker specific variability often do not perform well when other sources of acoustic variability are present. This has been observed to be the case for a class of approaches that performs speaker normalization through a process of spectral warping during feature analysis. One member of this class, often referred to as vocal tract length normalization (VTLN), estimates spectral warping factors by choosing the degree of warping that maximizes the average likelihood of a warped utterance with respect to the HMM model [1]. Another member of this class is an augmented state space acoustic decoder (referred to here as the MATE decoder), which uses a modified Viterbi algorithm to search for locally optimum degrees of spectral warping or temporal warping for *individual* frames [2]. We will show that the combination of discriminant feature space transformations (DFT) and noise robust processing techniques with frequency warping based speaker normalization can result in larger, more consistent, reductions in ASR word error rate (WER) over a wider range of acoustic degradations.

The performance of DFTs are also diminished when many sources of acoustic variability are present. We will demonstrate that reducing variability can also improve the notion of class separability in discriminant feature spaces and result in reduced ASR WER. Many DFT approaches have been proposed for reducing feature space dimensionality while maintaining the separability of phonetic classes in ASR. These approaches differ in the exact definition of the separability criterion and their assumptions about the distribution of the data within classes. Examples considered here include linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA) [3, 4]. The transformation parameters are estimated by maximizing a separability criterion that minimizes the sample covariance within a class relative to the separation of the sample means between classes. Additional variation resulting from environmental, channel, or speaker specific effects will only add to the relative within-class variation in the sample statistics and diminish effects on WER.

Performing linear discriminant analysis in a feature space that has been compensated to reduce any of these sources of variability should increase the effective class separability for the DFT. Towards this end, by performing discriminant analysis in an environment, channel, and speaker normalized feature space, we hope to demonstrate that larger performance increases can be achieved than can be achieved from any individual approach.

The paper is organized as follows. Section 2 contains a discussion of the methods used to compensate for speaker and environment variability. These include the MATE decoder and VTLN for speaker normalization and a set of techniques implemented as part of an ETSI robust advanced front end (AFE) for robust acoustic environment compensation [5]. Section 3 discusses the implementation of the DFT. The results of an experimental study evaluating the impact of combined feature space normalization and DFT on ASR performance is presented in Section 4. Finally, discussion and conclusions are presented in Section 5.

2. Speaker and Environment Compensation

This section provides a brief overview of the VTLN and MATE speaker normalization approaches and the AFE based robust environment and channel compensation techniques that are applied in this work. The interaction of these procedures with HDA/MLLT based discriminant feature spaces will be presented in Section 4.

2.1. VTLN based speaker normalization

This class of techniques selects from an ensemble of linear frequency warping functions, $\mathcal{G} = \{g^{\alpha_i}\}_{i=1}^{N}$ to produce a warped frequency scale, $f' = g^{\hat{\alpha}}(f)$. The optimum warping function, $g^{\hat{\alpha}}$ is chosen to maximize the average likelihood of a length T sequence of frequency warped cepstrum observation vectors, $X^{\alpha} = \{x_t^{\alpha}\}_{t=1}^{T}$, with respect to the HMM. For the experiments described in Section 4, there is a set of $N = 11$ possible linear warping functions equally spaced along a range from a minimum of twelve percent compression and a maximum of twelve percent expansion of the frequency axis.

VTLN can be implemented during recognition as a two pass

This work has been supported by NSERC

procedure. In the first pass, an initial hypothesized word string is generated. This initial word string is then used in a second pass to find the optimum $g^{\hat{\alpha}}$ by computing the likelihood of $i = 1, ..., N$ utterances, where each utterance is generated using warping function g^{α_i}, by performing a probabilistic alignment of X^{α_i} with the decoded word string.

2.2. Augmented State Space Acoustic Decoder

The MATE decoder, as developed by Miguel et al, is a modified Viterbi algorithm that is implemented in an augmented state space [2]. It allows frame-specific spectral warping functions to be estimated as part of the search for an optimum path. It was shown by Miguel that this same search procedure can also be used to estimate the optimum frame-specific time interval over which cepstrum difference coefficients are computed [2]. A description of this augmented state space will be provided here and the modified search algorithm will be briefly summarized.

A Viterbi beam search decoder for continuous speech recognition is implemented by propagating paths into the nodes of a two dimensional trellis. Each node of the trellis corresponds to one of M HMM states $\{q_j\}_{j=1}^{M}$ evaluated for observation vectors $x_t, t = 1, ..., L$. In the MATE decoder, the state space can potentially be expanded by a factor of N, where $N = N_\alpha$ is the size of the warping function ensemble described in Section 2.1. This effectively results in a three dimensional trellis. Each node of this augmented trellis corresponds to one of as many as $M' = N \cdot M$ states, $\{q_j^k\}_{j=1,k=1}^{M,N}$.

The states, $\{q_j^k\}_{k=1}^{N}$, in the existing implementation share the same observation densities as the state q_j in the original model for all $j = 1, ..., M$. This tying of the observation densities can be expressed as

$$b_j{}^k(x_t) = b_j(x_t),\ j = 1, ..., M,\ k = 1, ..., N, \qquad (1)$$

where $b_j()$ is the original Gaussian mixture observation density function for state j in the original model λ, x_t is a mel-frequency cepstrum observation vector at a frame t, and $b_j^k()$, is the augmented state space density function for state j and warping function k.

The optimum sequence of states is identified for the decoding process in a standard HMM using the Viterbi algorithm,

$$\phi_j(t) = \max_i \{\phi_i(t-1) \cdot a_{i,j}\} \cdot b_j(c_t). \qquad (2)$$

In Equation 2, $\phi_j(t)$ is the likelihood of the optimum path terminating in HMM state q_j at time t and $a_{i,j}$ is the transition probability from state q_i to state q_j. The max is computed over all states that are permitted by the HMM model to propagate into state q_j which, for a left-to-right HMM topology would be q_{j-1}.

In the MATE decoder, the optimum sequence of states in the augmented state space is identified using a modified Viterbi algorithm,

$$\phi_{j,n}(t) = \max_{i \in \mathcal{I}, \alpha_m \in \mathcal{A}} \left\{\phi_{i,m}(t-1) \cdot a_{i,j}^{m,n}\right\} \cdot b_j(c_t^{\alpha_n}). \quad (3)$$

In Equation 3, $\phi_{j,n}(t)$ is the likelihood of the optimum path terminating in state q_j^n at time t and $a_{i,j}^{m,n}$ is the transition probability from state q_i^m to state q_j^n. The max is computed over all states that are permitted by the HMM model to propagate into state q_j^n.

Structural constraints can be placed on the transformations, g^{α_n}, that are permitted at state q_i^n in the augmented state space. These constraints can be applied by setting a subset of the transition probabilities, $a_{i,j}^{m,n}$ equal to zero. Transition probabilities were constrained so that the frequency warping transformations applied to adjacent frames were required to be taken from adjacent indices in the ensemble $\mathcal{G}$. This implies that $a_{i,j}^{m,n} = 0$ if $|m - n| > 1$. These constraints have the effect of reducing the computational complexity in search. Furthermore, they also provide a means for limiting the degrees of freedom in the application of spectral transformations to reflect a more physiologically plausible degree of variability.

2.3. Robust Feature Analysis

All of the experiments reported in Section 4 are performed using mel-frequency cepstrum coefficient (MFCC) feature analysis. First and second order cepstrum difference coefficients are computed to model temporal dynamics in speech. However, there is no mechanism in this standard configuration for compensating with respect to acoustic environment or channel variability. To investigate how feature compensation algorithms can impact the performance of speaker normalization and discriminant feature transformations, a noise-robust advanced DSR front end (AFE) was applied [5].

There are three major robust processing steps that are applied in the AFE. First, a two stage Wiener filter is applied. Second, an SNR-dependent waveform processing procedure is performed which estimates a weighting function to emphasize the high SNR portions of the waveform and de-emphasize the low SNR portions of the waveform [5]. This combination of noise reduction and SNR dependent waveform processing serves to increase the effective SNR prior to computation of the MFCC thereby reducing the impact of additive environmental noise. Finally, after cepstrum computation, blind equalization is performed to reduce the effects of convolutional distortion that may arise from transducer or channel dependent mismatch.

3. Discriminant Feature Transformations

This section provides a brief summary of discriminant feature transformations (DFTs). The DFT is applied to estimating a feature space transformation from a high dimensional feature space to a lower dimensional space while maximizing the separability betwen classes in the target feature space [3].

Assume that we are given a set of d dimensional data vectors $X = \{x_t\}_{t=1}^{T}$, each belonging to one of a set of L classes $C = \{c_l\}_{l=1}^{L}$ where each class contains N_l vectors. For each class, we have the sample means and covariances

$$\mu_l = \frac{1}{N_l} \sum_{x_t \in c_l} x_t \qquad \Sigma_l = \frac{1}{N_l} \sum_{x_t \in c_l} (x_t - \mu_j)(x_t - \mu_j)^T$$

The optimization criteria for LDA are based on the within class scatter matrix, and the between class scatter matrix, S_B,

$$S_W = \frac{1}{N} \sum_{l=1}^{L} N_l \Sigma_l \qquad S_B = \frac{1}{N} \sum_{l=1}^{L} N_l (\mu_l - \mu)(\mu_l - \mu)^T$$

where $N = \sum_{l=1}^{L} N_l$ and $\mu = \frac{1}{N} \sum_{l=1}^{L} \mu_l$.

Our goal is to estimate a linear transformation from the d dimensional x to a p dimensional y where $p < d$. This transformation takes the form of a $p \times d$ dimensional matrix A with linearly independent rows and columns such that $y = Ax$. In LDA, the parameters of A can be estimated to optimize a measure of class separability in a transformed space which is

based on the ratio between the between class scatter to the within class scatter [3]:

$$J_{LDA}(A) = \log \left| (AS_W A^T)^{-1} (AS_B A) \right| \quad (4)$$

HDA extends this criterion to a more general separability criterion in the transformed space that maximizes the separability only in the projected dimension [4]:

$$J_{HDA}(A) = \sum_{l=1}^{L} N \log \left| AS_B A^T \right| - N_j \log \left| A\Sigma_l A^T \right| \quad (5)$$

Since the data within each class for both LDA and HDA are assumed to have full covariance distributions, there is a mis-match to the diagonal covariance assumption that is used for observation densities in an HMM. To compensate for this mismatch, another discriminant transformation can be applied in the p dimensional transformed space that enforces the diagonal covariance constraint in the transformed space. This is referred to as maximum likelihood linear transformation (MLLT) since it produces a discriminant feature space that maximizes the likelihood of the data under a diagonal class covariance constraint [4].

In our implementation, an HDA transformation is estimated for an input feature space where the data vectors correspond to the concatenation of nine cepstrum feature vectors. With thirteen component cepstrum, this corresponds to a $d = 117$ dimensional space. The dimensionality of the transformed feature space was chosen to be $p = 39$. A 39×39 component MLLT transformation was estimated and applied to the HDA transformed features. The classes for the DFT were defined to be the individual states of the HMM. Working on a limited vocabulary connected digit recognition domain, there were a total of approximately 180 HMM states resulting in the same number of classes in the discriminant space.

4. Experimental Study

This section addresses the interaction between speaker normalization, environment compensation, and DFT in terms of their impact on ASR performance. Experimental comparisons were performed for two ASR task domains where the first domain was based on simulated acoustic environments and the other was based on utterances collected in a range of actual automobile driving scenarios. After describing the databases associated with these task domains and the baseline HMM ASR system configuration, three experimental comparisons are presented. First, the effect of environmental variability on the ability of speaker normalization to reduce ASR WER is described. Second, the ability of both robust feature analysis and DFT to improve the performance of VTLN and MATE speaker normalization procedures in difficult acoustic environments is described. Third, the effect of training the DFT parameters using speaker normalized data on WER is considered.

4.1. Speech Corpora and ASR Platform

There were two different speech corpora that were used to evaluate the performance of the speaker normalization and discriminant feature transformation (DFT) scenarios discussed in Sections 2 and 3. Both corpora consisted of utterances of connected digits. The first task domain corresponds to a subset of the Aurora 2 database that included four different simulated acoustic environments. These were obtained by creating recordings of subway, speech babble, automobile, and exhibition hall conditions and adding them to speech utterances recorded in clean conditions under different signal-to-noise (SNR) ratio assumptions. A total of 8440 utterances (27727 digits) were used for training and 4004 utterances (13159) digits were used for evaluation. All of the ASR results presented in Tables 1 and 2 were obtained using HMM models that were trained from multiple conditions and multiple SNRs.

The second task domain was the Aurora 3 subset of the Spanish Language SpeechDat Car database. This corpus was collected using hands-free and close-talking microphones from a population of 160 speakers operating a vehicle under several driving conditions. These conditions included a stopped car with motor running, driving in traffic at low speeds, and high speed driving. A total of 3292 of these utterances (18334 digits) were used for training acoustic HMM models and 1522 utterances (5012 digits) were used for testing. All of the ASR results presented in Table 3 were obtained using the above scenario.

The baseline ASR system is the same as that described in a previous study of VTLN and MATE based speaker normalization [2]. HMM word models with 16 states and 3 Gaussian densities per state were used to represent Spanish digit utterances. The baseline system uses ETSI standard MFCC feature analysis [6]. All procedures for training HMM models for the MATE decoder are also the same as reported in [2].

4.2. Experimental Results

Table 1 demonstrates how the effects of VTLN and MATE speaker normalization are significantly diminished in the presence of external acoustic variability. The table displays the WER for the baseline ASR system described above, the baseline system implemented with VTLN based speaker normalization as described in Section 2.1, and the MATE based decoder described in Section 2.2. All three systems are trained under the multi-condition scenario described above and evaluated under clean high SNR and noisy 15 dB SNR conditions on the Aurora 2 database. All of the systems in Table 1 used the ETSI standard MFCC feature analysis with no environment or channel normalization techniques applied. It is clear from Table 1 that both VTLN and MATE based

Aurora 2 Word Error Rate (Improvement)		
System	Clean	15 dB SNR
Baseline	1.40%	2.18%
VTLN	1.21% (13.6%)	2.08% (4.53%)
MATE	1.01% (27.8%)	1.94% (11.15%)

Table 1: WER for VTLN and MATE using MFCC features.

speaker normalization provide significant WER reduction with respect to the baseline system under clean testing conditions. The relative improvements are shown in the table to be 13.6 percent and 27.8 respectively. However, it is also clear that these relative improvements are reduced by approximately a factor of three under moderately noisy conditions.

In order to quantify the effects of reduced environmental variability on speaker normalization performance, the three systems shown in Table 1 were evaluated under the same noisy conditions using the robust AFE feature analysis. The first column of Table 2 displays the WER for the Baseline, VTLN, and MATE systems when the AFE feature analysis is used for all these systems under noisy 15 dB SNR test

conditions. Comparing the relative reduction in WER for VTLN and MATE shown in the first column of Table 2 to the relative improvements for those systems shown in the second column of Table 1, it is clear that the use of AFE results in considerably greater improvements being introduced by MATE and VTLN speaker normalization.

Aurora 2 Word Error Rate (Improvement)			
System	AFE	DFT	DFT+SNF
Baseline	1.95%	1.66%	
VTLN	1.80% (7.7%)	1.57% (5.4%)	1.50% (4.4%)
MATE	1.60% (17.9%)	1.42% (15.1%)	1.40% (1.6%)

Table 2: WER for VTLN and MATE using AFE features, DFT using AFE, and DFT using speaker norm. features (SNF).

The combined effect of applying both robust feature analysis for reducing environment and channel variability and discriminant feature transformations to improve class separability can be observed by comparing the first and second columns of Table 2. The HDA and MLLT parameters are estimated directly from the features generated by the AFE based feature analysis and then used to transform features during HMM training and recognition. It is clear from the results shown in Table 2 that the DFT produces significant WER reduction for all three systems.

The effect of reducing speaker variability prior to performing linear discriminant analysis is given in last column of Table 2. For VTLN, each training utterance was warped to maximize $P(X^{\alpha}|\lambda)$ and the warped utterances were used for estimating the HDA/MLLT matrices. For MATE, each utterance was warped using frame-specific warping factors selected by the MATE decoder. These warped utterances were then used for training the HDA/MLLT matrices. It is shown in Table 2 that estimating HDA/MLLT matrices from speaker normalized data has a small effect on WER reduction for this noisy Aurora 2 data.

It can be shown that the value of the LDA optimization criterion given in Equation 4 is proportional to the sum of the magnitudes of the p largest eigenvalues of $S_W^{-1}S_B$. To give an indication of the degree to which the robust AFE feature analysis can reduce the within class variability and influence the separability criterion used for LDA, the 30 largest eigenvalues are displayed in Figure 1 for two input feature spaces. Both examples are computed from simulated noisy observation data obtained from the Aurora 2 database described in Section 4.1. The bottom curve in the figure corresponds to eigenvalues that were obtained from LDA estimated from standard MFCC features with no robust processing applied. The top curve corresponds to eignenvalues obtained from LDA estimated from features estimated using robust AFE feature analysis. From the differences between the two curves, it appears that class separability and presumably also ASR performance would be greatly enhanced by using the robust AFE. This is indeed supported by the observed WER reduction using the baseline ASR system when applying DFT to standard MFCC and AFE features on the Aurora 2 database. A 2.2% WER reduction was obtained by DFT using MFCC features and a much larger reduction of 14.9

Table 3 shows the results of the evaluation of the speaker normalization and DFT procedures on the Spanish language Aurora 3 database. This database differs from the Aurora 2 database in that it represents actual rather than simulated acoustic environments. However, it also contains approximately one third of the number of utterances for training and testing that are contained in the Aurora 2 database. As a result, the VTLN and MATE procedures, which are not heavily reliant on training, show consistent improvement. However, when used with the DFT, which does rely on having sufficient data for training HDA/MLLT transformations, there is a less consistent reduction in WER.

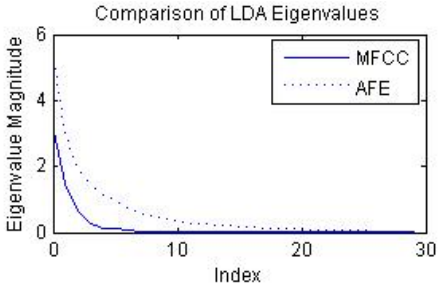

Figure 1: LDA eigenvals for MFCC and AFE features

Aurora 3 Word Error Rate (Improvement		
System	AFE	AFE+DFT
Baseline	2.0%	1.8%
VTLN	1.8% (10.0%)	1.6% (11.1%)
MATE	1.7% (15.0%)	1.7% (5.5%)

Table 3: WER on in-car database for VTLN and MATE using AFE features and DFT.

5. Conclusions

This paper has demonstrated the WER reductions that can be obtained through the combined application of speaker normalization, AFE feature analysis, and DFT on noise corrupted connected digit tasks. The MATE decoder provided an 18% WER reduction over the baseline system when applied to features compensated by the AFE. An additional 12% WER reduction was obtained when MATE was applied in the discriminant feature space, and a minor reduction in WER was obtained when the DFT was in turn trained with and applied to MATE normalized features.

6. References

[1] L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans on Speech and Audio Processing*, vol. 6, January 1998.

[2] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," *Proc. Int. Conf. on Spoken Lang. Proc.*, Sept. 2005.

[3] K. Fukunaga, *Statistical pattern recognition*, 2nd ed. Boston: Academic Press, 1990.

[4] G. Saon, M. Padmanabhan R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *Proc. Int. Conf. Acous., Speech, and Sig. Proc.*, May 2000.

[5] D. Macho, L. Mauuary, B. Noi, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," *Proc. Int. Conf. on Spoken Language Proc.*, Oct. 2002.

[6] D. Pearce, "An overview of the ETSI standards activities for distributed speech recognition front-ends," *Applied Voice Input/Output Society Conference AVIOS2000*, May 2000.