ELIMINATING INTER-SPEAKER VARIABILITY PRIOR TO DISCRIMINANT TRANSFORMS

George Saon, Mukund Padmanabhan, Ramesh Gopinath

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
E-mail: {gsaon,mukund rameshg}@us.ibm.com, Phone: (914)-945-2985

ABSTRACT

This paper shows the impact of speaker normalization techniques such as vocal tract length normalization (VTLN) and speaker-adaptive training (SAT) prior to discriminant feature space transforms, such as LDA. We demonstrate that removing the inter-speaker variability by using speaker compensation methods results in improved discrimination as measured by the LDA eigenvalues and also in improved classification accuracy (as measured by the word error rate). Experimental results on the SPINE (speech in noisy environments) database indicate an improvement of up to 5% relative over the standard case where speaker adaptation (during testing and training) is applied after the LDA transform which is trained in a speaker independent manner.

We conjecture that performing linear discriminant analysis in a canonical feature space (or speaker normalized space) is more effective than LDA in a speaker independent space because the eigenvectors will carve a subspace of maximum intra-speaker phonetic separability whereas in the latter case this subspace is also defined by the inter-speaker variability. Indeed, we will show that the more normalization is performed (first VTLN, then SAT) the higher the LDA eigenvalues become.

1. INTRODUCTION

One of the most commonly used feature extraction methods in speech recognition consists in first computing a fixed number of cepstral coefficients for every speech frame (usually 13) and then augmenting the feature vector with dynamic information from the adjacent frames. On the one hand, the cepstral feature extraction process is well motivated by mimicing the operations that are carried out in the human auditory system, however, there is no explicit attempt to discriminate between the different phonetic classes. On the other hand, one could argue that as the fundamental problem in speech recognition is to discriminate between phonetic classes on the basis of the observed feature vector, the feature extraction process should be designed so as to achieve this goal. One way to incorporate both of these goals is to apply a discriminant linear projection on the extracted cepstra that helps to discriminate between the phonetic classes.

Linear discriminant analysis [1, 2] is a standard technique in statistical pattern classification for dimensionality reduction with a minimal loss in discrimination. Straightforward implementations of LDA to project the cepstral features into a discriminant feature space have however had only a limited amount of success because the transformations that are produced by LDA are often inconsistent with the assumptions made in speech recognition systems. Chief among these is the assumption that the pdf of the LDA transformed features can be modeled with diagonal covariance gaussians for the different classes. Consequently, recent work [9, 3, 4] has focused on designing the projection such that this assumption is satisfied. For instance, when this assumption is directly incorporated in a discriminant objective function [9], it leads to a ”maximum likelihood discriminant” projection that provides significant improvements over the baseline system.

All of this prior work, however, is characterized by the fact that the projection is computed to perform dimensionality reduction at the first (speaker-independent) stage of the processing. Now, the main objective of discriminant projections is to minimize the variation of the projected feature within a particular class, while maximizing the distance between the projected means of the different classes. As the training data for a speaker independent system is comprised of speech from a number of different speakers, the variation of the projected features within a particular class has an inherent component as well as an inter-speaker component. For the purposes of discriminating between phonetic classes, we are really interested only in focusing on the inherent variation, rather than the inter-speaker variation. In order to achieve our objective, we can take the help of speaker adaptation techniques [7, 8] that are focused on improving the performance of speech recognition systems by ”canonicalizing” the feature space i.e. by eliminating as much of the inter-speaker variability as possible. This is in effect equivalent to first ”canonicalizing” the feature space with some speaker normalizing scheme and then computing a discriminant transform that separates the phonetic classes out in the canonicalized space.

In this paper, we demonstrate that removing the inter-speaker variability by using speaker compensation methods results in improved discrimination as measured by the LDA eigenvalues and also in improved classification accuracy (as measured by the word error rate). Experimental results on the SPINE (speech in noisy environments) database indicate an improvement of up to 5% relative over the standard case where speaker adaptation (during testing and training) is applied after the LDA transform which is trained in a speaker independent manner.

The paper is organized as follows: in section 2 we describe the formulation for intra and inter-speaker LDA and show that the leading eigenvalues of LDA improve when the feature space is canonicalized by speaker normalization. In Section 3 we describe the experiments and results and section 4 provides a final discussion.

2. INTRA AND INTER-SPEAKER LDA

Let the labeled acoustic training data be $\{(x_i, c_i, s_i)\}_{1\leq i\leq N}$, where $x_i \in \mathbb{R}^n$, $c_i \in \mathcal{C}$ and $s_i \in \mathcal{S}$ are respectively the acoustic vector, class id and speaker id associated with the i^{th} training sample. Then, the counts, sample means and covariances associated with the class-speaker pair (c, s) are

$$N_{c,s} = \sum_{(c_i,s_i)=(c,s)} 1, \quad \mu_{c,s} = \frac{1}{N_{c,s}} \sum_{(c_i,s_i)=(c,s)} x_i,$$

$$\Sigma_{c,s} = \frac{1}{N_{c,s}} \sum_{(c_i,s_i)=(c,s)} (x_i - \mu_{c,s})(x_i - \mu_{c,s})^T$$

and therefore the counts, sample means and covariances for class c are

$$N_c = \sum_s N_{c,s}, \quad \mu_c = \frac{1}{N_c} \sum_s N_{c,s}\mu_{c,s},$$

$$\Sigma_c = B_c^{\mathcal{S}} + \frac{1}{N_c} \sum_s N_{c,s}\Sigma_{c,s},$$

where $B_c^{\mathcal{S}}$ is the between-speaker covariance matrix for class c:

$$B_c^{\mathcal{S}} = \frac{1}{N_c} \sum_s N_{c,s}(\mu_{c,s} - \mu_c)(\mu_{c,s} - \mu_c)^T$$

Now the within-class covariance W of the data is given by

$$\begin{aligned} W &= \frac{1}{N} \sum_c N_c\Sigma_c = \frac{1}{N} \left[\sum_{c,s} N_{c,s}\Sigma_{c,s} + \sum_c N_c B_c^{\mathcal{S}}\right] \\ &= \frac{1}{N} \sum_{c,s} N_{c,s}\Sigma_{c,s} + B^{\mathcal{S}} \qquad (1) \end{aligned}$$

and hence the total covariance is given by

$$T = W + B^{\mathcal{C}} = \frac{1}{N} \sum_{c,s} N_{c,s}\Sigma_{c,s} + B^{\mathcal{S}} + B^{\mathcal{C}} \qquad (2)$$

In standard Linear Discriminant Analysis (LDA) one finds a projection matrix $\theta \in \mathbb{R}^{p\times n}$ of full row rank such that the ratio of determinants of the projected total covariance and the projected within-class covariance is maximized, i.e.:

$$\theta^{\text{LDA}} = \underset{\theta\in\mathbb{R}^{p\times n}}{\text{argmax}} \frac{|\theta T\theta^T|}{|\theta W\theta^T|} \qquad (3)$$

The solution to (3) can be found by solving the generalized eigenvalue problem: $Tx = \lambda Wx$ or equivalently, by finding the p largest eigenvalues of $W^{-1}T$ (assuming W is not rank defficient). The transposed eigenvectors corresponding to these eigenvalues will form the rows of the projection matrix θ. The maximum value of the objective function (2) corresponds to the product of the p largest eigenvalues and for $p = n$ this product is equal to the ratio of the determinants of T and W.

Our proposed algorithm is to compute the LDA after a speaker-normalization transformation (like SAT or VTL). Analytically studying this effect for SAT or VTL seems very difficult. Nevertheless, to gain insight into what the algorithm does we study what happens when the speaker-normalization transformation is *ideal*. In the ideal case it is reasonable ot assume that the *normalized-data* is such that the all the speaker means for a given class are all identical, i.e., $\mu_{c,s} = \mu_c$ and therefore $B_c^{\mathcal{S}} = 0, \forall c \in C$. The within-class and total covariance of the normalized data are therefore (from Eqn. 1)

$$W^N = \frac{1}{N} \sum_c N_c\Sigma_c = \frac{1}{N} \sum_{c,s} N_{c,s}\Sigma_{c,s}, \quad T^N = W^N + B^{\mathcal{C}} \qquad (4)$$

What is the relationship between the LDA's before and after ideal speaker-normalization? That is, what is the relationship between the following two ratios:

- Standard LDA: $\frac{|\theta T\theta^T|}{|\theta W\theta^T|}$ and
- Speaker-Normalized LDA: $\frac{|\theta T^N\theta^T|}{|\theta W^N\theta^T|}$

The main result of this paper is the following inequality (that follows from Proposition 1 below):

$$\frac{|T|}{|W|} = \frac{|W + B^{\mathcal{C}}|}{|W|} = \frac{|W^N + B^{\mathcal{S}} + B^{\mathcal{C}}|}{|W^N + B^{\mathcal{S}}|} < \frac{|W^N + B^{\mathcal{C}}|}{|W^N|} = \frac{|T^N|}{|W^N|} \qquad (5)$$

which implies that, for a full rank transformation, the objective function for normalized LDA is always higher than that for standard LDA.

Let $\mathcal{S}$ the set of all s.p.d. matrices of order n, that is: $\mathcal{S} = \{A \in \mathbb{R}^{n\times n} | A = A^T, x^T Ax \geq 0, \forall x \in \mathbb{R}^n\}$. Define the binary relation "$\succ$" on $\mathcal{S} \times \mathcal{S}$ by $A \succ B$ if $A - B \in \mathcal{S}$, that is, if their difference is an s.p.d. matrix. Two simple lemmas are immediate.

Lemma 1 If $A \succ B$ then for any non-singular matrix C (not necessarily s.p.d.) (i) $A + C \succ B + C$, (ii) $C^T AC \succ C^T BC$ and (iii) $B^{-1} \succ A^{-1}$.

Proof: As (i) is evident, we will only prove (ii) and (iii). For (ii), we have $x^T C^T (A-B)Cx = (Cx)^T (A-B)Cx = y^T (A-B)y \geq 0$ because $A-B$ is s.p.d. For (iii), the proof follows from the simultaneous diagonalization of A and B (as in Lemma 2) and from (ii). □

Lemma 2 *If $A \succ B$ then $|A| > |B|$.*

Proof: Indeed, $C := A - B$ is s.p.d. and according to [6] p. 312, case 3, there exists a non-singular P such that $B = PP^T$ and $C = PDP^T$ where $D = diag(\lambda_1, \ldots, \lambda_n)$. Then $|A| > |B| \Leftrightarrow |D + I| > |I|$ which is true because of the positivity of the λ_i's. □

Proposition 1 *If $A, B, C \in \mathbb{R}^{n\times n}$ are three symmetric positive definite (s.p.d.) matrices then the following inequality holds:*

$$\frac{|A+B+C|}{|A+C|} < \frac{|A+B|}{|A|} \tag{6}$$

Proof:

$$\begin{aligned}
&\frac{|A+B+C|}{|A+C|} < \frac{|A+B|}{|A|}\\
\Leftrightarrow\ & |(A+C)^{-1}(A+B+C)| < |A^{-1}(A+B)|\\
\Leftrightarrow\ & |I + (A+C)^{-1}B| < |I + A^{-1}B|\\
\Leftrightarrow\ & |B^{\frac{1}{2}}||I + (A+C)^{-1}B||B^{-\frac{1}{2}}| < |B^{\frac{1}{2}}||I + A^{-1}B||B^{-\frac{1}{2}}|\\
\Leftrightarrow\ & |I + B^{\frac{1}{2}}(A+C)^{-1}B^{\frac{1}{2}}| < |I + B^{\frac{1}{2}}A^{-1}B^{\frac{1}{2}}|
\end{aligned} \tag{7}$$

Now, $A + C \succ A$ implies $A^{-1} \succ (A+C)^{-1}$ from lemma 1 (iii). Therefore, $B^{\frac{1}{2}}A^{-1}B^{\frac{1}{2}} \succ B^{\frac{1}{2}}(A+C)^{-1}B^{\frac{1}{2}}$ from lemma 1 (ii). Finally, $I + B^{\frac{1}{2}}A^{-1}B^{\frac{1}{2}} \succ I + B^{\frac{1}{2}}(A + C)^{-1}B^{\frac{1}{2}}$ from lemma 1 (i). The proof follows from applying lemma 2 to the resulting matrices. □

3. EXPERIMENTS AND RESULTS

The speech recognition experiments were conducted on the SPINE (speech in noisy environments) database. The speech data consists of conversations between two communicators working on a collaborative, Battleship-like task in which they seek and shoot at targets. The SPINE training data consists of 12 hours of conversation, including 20 speakers (10 speaker pairs) in four noise environments (quiet, office, humvee, and air craft carrier). The vocabulary size and perplexity are fairly low (1.2K words and 18 respectively) but the difficulty of the task comes from the various simulated noise environments.

Speech is coded into 25 ms frames, with a frame-shift of 10 ms. Each frame is represented by a feature vector of 13 Mel frequency-warped cepstral coefficients (MFCC) computed from a 24-filter Mel filterbank spanning the 0 Hz - 4.0 kHz frequency range. Every 9 consecutive cepstral frames are spliced together and projected down to 60 dimensions using linear discriminant analysis (LDA). The range of this transformation is further diagonalized by means of a maximum likelihood linear transform (MLLT) [3].

We have experimented with various systems of similar size both in terms of the number of context-dependent HMM states (roughly 1800) and in the number of diagonal Gaussian mixture components (around 25K 60-dimensional Gaussians). The systems differ depending on where the LDA projection is applied. For the first system, LDA is applied on the speaker-independent MFCC frames. For the second system, the transform is applied on vocal tract length normalized features and for the third system an additional speaker-adaptive training step is performed after VTLN but *before* LDA. In other words, for the third system, we model 117 dimensional features obtained by splicing every 9 consecutive 13 dimensional VTLN-warped cepstral frames. This space is first diagonalized by means of a 117 × 117 MLLT transform. We then seed the canonical model for SAT with 11K 117-dimensional Gaussians obtained by clustering these features. Next, we compute a feature space transform [5, 10] for the training data for each speaker, which maximizes the likelihood of the transformed data given the canonical model. We then estimate a 60 × 117 LDA transform on the (per speaker) linearly warped, VTLN-warped features and an MLLT transform in the resulting 60 dimensional space. The effect of these various speaker normalization steps on the LDA eigenvalues is illustrated in figure 1. This appears to be consistent with the results of the previous section.

The test set consists of the first 18 converations (36 speakers) of last year's evaluation set and has 1.5K utterances (8.3K words). We will report detailed results only for the last two systems (LDA applied after VTLN and LDA applied after VTLN+SAT) as these systems gave the best performance. The rows of table 1 correspond to these two

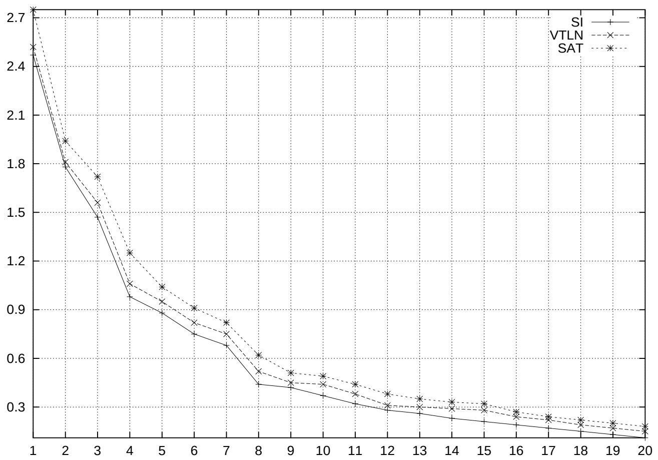

Figure 1: The 20 largest LDA eigenvalues for the SI, VTLN and SAT feature spaces.

Transform	No adaptation	AD1	AD2
VTLN-LDA	28.0%	22.7%	22.1%
VTLN-SAT-LDA	28.8%	22.5%	21.1%

Table 1: Word error rates for the various systems/passes.

systems and the columns to the various adaptation steps. No adaptation means the baseline performance after VTLN. The first adaptation step consists in computing a 60 and respectively, a 117 dimensional feature space MLLR transform. For the SAT+LDA case, we have the luxury of also adapting the 60 dimensional features after SAT-LDA which we referred to as the second adaptation step. In order to make a fair comparison with SAT-LDA, we perform the 60-dimensional adaptation step twice for VTLN-LDA (using the updated decoding hypothesis from the previous adaptation pass).

4. CONCLUSION

Analyzing the results from the previous table, several conclusions can be drawn: the baseline performance of VTLN-LDA is better than that of SAT-LDA because, for the latter, there is a mismatch between the canonical space in which the transform was trained and the actual (untransformed) adaptation data. However, after the first MLLR transform, SAT-LDA appears to be slightly more effective than VTLN-LDA although the statistical significance of the difference can be questioned. The main difference between the two is apparent after the last adaptation step where iterating the feature space adaptation transform for VTLN-LDA does result in smaller gains than for SAT-LDA. This may be explained by the fact that the second transform for SAT-LDA makes use of the full-resolution 60-dimensional model (25K Gaussians) whereas the first (117-dimensional) transform needed a much more coarser canonical model (11K Gaussians) so there may be a complementarity effect between the two transforms.

5. REFERENCES

[1] R. O. Duda and P. B. Hart. Pattern classification and scene analysis. *Wiley*, New York, 1973.

[2] K. Fukunaga. Introduction to statistical pattern recognition. *Academic Press*, New York, 1990.

[3] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. *Proceedings of ICASSP'98*, Denver, 1998.

[4] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281, 1999.

[5] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1997.

[6] S. R. Searle. Matrix algebra useful for statistics. *Wiley Series in Probability and Mathematical Statistics*, New York, 1982.

[7] S. Wegman, D. McAllaster, J. Orloff, and B. Peskin, ”Speaker normalization on conversational telephone speech”, *Proceedings of ICASSP'96*, 1996.

[8] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul, ”A compact model for speaker-adaptive training”, *Proceedings of ICASSP'96*, 1996.

[9] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, ”Maximum Likelihood Discriminant Feature Spaces”, *Proceedings of ICASSP'00*, Istanbul, 2000.

[10] G. Saon, G. Zweig and M. Padmanabhan. Linear feature space projections for speaker adaptation. *Proceedings of ICASSP 2001*, Salt Lake City, 2001.