Analysis of information in speech and its application in speech recognition

Sachin S. Kajarekar[1] and Hynek Hermansky[1,2]

[1] Oregon Graduate Institute of Science and Technology,
Beaverton, OR, USA.
{sachin,hynek}@ece.ogi.edu
http://www.asp.ece.ogi.edu/index.html
[2] International Computer Science Institute,
Berkeley, CA, USA.

Abstract. Previous work analyzed the information in speech using analysis of variance (ANOVA). ANOVA assumes that sources of information (phone, speaker, and channel) are univariate gaussian. The sources of information, however, are not unimodal gaussian. Phones in speech recognition, e.g., are generally modeled using a multi-state, multi-mixture model. Therefore, this work extends ANOVA by assuming phones with 3 state, single mixture distribution and 5 state, single mixture distribution. This multi-state model was obtained by extracting variability due to position within phone from the error term in ANOVA. Further, linear discriminant analysis (LDA) is used to design discriminant features that better represent both the phone-induced variability and the position-within-phone variability. These features perform significantly better than conventional discriminant features obtained from 1-state phone model on continuous digit recognition task.

1 Introduction

Speech signal and features extracted from the signal contain various sources of information. For speech and speaker recognition, phone, speaker, and communication channel (henceforth referred as channel) are the most important sources. Previous work [2, 3], analyzed information in speech using analysis of variance (ANOVA) [8]. The contribution of phone, speaker and channel was estimated in spectral and temporal domains.

As explained in section 2, ANOVA assumes that sources have unimodal gaussian distribution[1]. Sources of information in speech however are more complex, e.g., phones are typically modeled using a multi-state, multi-mixture model. In this paper, therefore, we extend ANOVA to measure phone information assuming such multi-state gaussian distribution. We show that sources can be recombined into useful and harmful sources. Thus, e.g. for speaker independent automatic recognition of speech (ASR), phone information represents the useful information and all other sources contribute harmful information. This relates ANOVA

[1] Information is proportional to variance if sources have gaussian distribution [7].

to our work on discriminant features (e.g., FIR RASTA filters in temporal domain [1] and discriminant bases in spectral domain [5]) using linear discriminant analysis (LDA). We show that improvement in the estimate of phone information could lead to more robust features.

The paper is organized as follows. In section 2, we give a brief overview of ANOVA and extend it for multi-state sources. This is followed by a short description of LDA. In section 3, experimental setup is described and in section 4, we present results of ANOVA and LDA. We conclude with summary of our work in section 5.

2 ANOVA and LDA

2.1 ANOVA

ANOVA assumes that sources have unimodal gaussian distributions. From given data, sources are segregated as follows. First, mean and variance of each source is estimated and any observation X_{ijkl} is expressed as a linear combination of means of sources – phone (i), speaker (j), channel (k), their interactions and an error term

$$X_{ijk} = \bar{X}_{....} + \bar{X}_{i...} + \bar{X}_{.j..} + \bar{X}_{..k.} + \bar{X}_{ij..} + \bar{X}_{.jk.} + \bar{X}_{i.k.} + \bar{X}_{ijk.} + \epsilon_{ijkl}, \quad (1)$$

where $\bar{X}$ represents the mean of the source or source interaction. Then the decomposition of total information (Σ_{total}) follows from above definition as

$$\Sigma_{total} = \Sigma_i + \Sigma_j + \Sigma_k + \Sigma_{ij} + \Sigma_{jk} + \Sigma_{ik} + \Sigma_{error}. \quad (2)$$

For speech recognition, we are interested only in the phone information. So we can rewrite the above equation as

$$\begin{aligned} \Sigma_{total} &= \Sigma_p + \tilde{\Sigma}_{error} \\ &= \frac{1}{a}\sum_{i=1}^{a}(\bar{X}_{i...} - \bar{X}_{....})^2 + \frac{1}{abcd}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{d}(X_{ijkl} - \bar{X}_{i...})^2. \end{aligned} \quad (3)$$

So far, we assumed a single-state model for phone. For multiple-state phone model, the above equation becomes

$$\Sigma_{total} = \frac{1}{aS}\sum_{i=1}^{a}\sum_{s=1}^{S}(\bar{X}^s_{i...} - \bar{X}_{....})^2 + \frac{1}{abcdS}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{d}(X^s_{ijkl} - \bar{X}^s_{i...})^2 \quad (4)$$

where S =number of states. In the above equation, first term is rewritten as

$$\frac{1}{aS}\sum_{i=1}^{a}\sum_{s=1}^{S}(\bar{X}^s_{i...} - \bar{X}_{....})^2 = \frac{1}{aS}\sum_{i=1}^{a}\sum_{s=1}^{S}(\bar{X}^s_{i...} - \bar{X}_{i...})^2 + \frac{1}{a}\sum_{i=1}^{a}(\bar{X}_{i...} - \bar{X}_{....})^2. \quad (5)$$

This shows that the phone information (assuming a multiple state model) is the sum of the phone information with a single-state model (second term) and the average position within-phone information (first term).

2.2 LDA

Given that features represent phone classes, LDA [6] is used to select directions that represent the maximum useful (across-class, Σ_a) information while suppressing harmful (within-class, Σ_w) information. We refer to these directions as linear discriminants. For speech recognition, phone information is the useful information and other sources represent the harmful information. Therefore, $\Sigma_a = \Sigma_p$ and $\Sigma_w = \tilde{\Sigma}_{error}$. The discriminant directions (E) are obtained from these matrices as

$$E = eig(\Sigma_{error}^{-1} * \Sigma_p). \quad (6)$$

Discriminant features ($\tilde{X}$) are obtained by projecting original features (X) on discriminant directions (E), i.e., $\tilde{X} = XE$

3 Experimental Setup

OGI Stories [4] database was analyzed with ANOVA. The database contains about 3 hours of phonetically hand-labeled conversational speech. That represents 210 speakers, speaking for about 50 sec each, though different telephone channels. A set of most frequently occurring 38 phonemes from this database was used in this study.

15 critical band spectra, calculated using 25 ms hamming window at 100 Hz, were used as features in the spectral analysis. For the temporal analysis, each band was analyzed independently. A 101 dimensional feature vectors [1] were used which were labeled by phonemes in their center.

Discriminant were derived independently in spectral and temporal domains using features described above. Spectral discriminants were referred as spectral bases [5] and temporal discriminants were referred as FIR RASTA filters [1]. Discriminants were also derived from sequential optimization in frequency and time domains as follows. First, spectral discriminant bases were derived. The critical band spectra were projected on 8 spectral discriminants to form new 8 dimensional feature vectors. Temporal discriminants were designed separately for temporal streams of these 8 features. These temporal streams were filtered using first 3 temporal discriminants from each stream to obtain $8 * 3 = 24$ features.

The discriminant features were evaluated on the continuous-digit recognition task. Digits were modeled using 23 context independent monophone classes that were trained using 5-state, 3-mixture HMM. The baseline system used 24 features – 8 cepstral coefficients and their derivatives (Δs) and double derivatives ($\Delta\Delta$s)in time. Temporal means were removed from features in each file (cepstral mean normalization). Spectral LDA (sLDA) system used 8 spectral discriminant features + 8 Δs + 8 $\Delta\Delta$s with mean normalization. Temporal LDA (tLDA) system used 45 temporal discriminant features. For system using joint discriminants (sLDA+tLDA), it was empirically observed that temporal discriminants from first stream of spectral discriminants were less noisy than those obtained from subsequent spectral streams. We compared 1) the system with features derived using identical discriminants (derived from the first stream) for all 8 streams to 2) the system with features derived using different temporal discriminants for each stream. Significantly better performance was obtained using the system with identical discriminants. Results of such a system are reported in this paper.

4 Results

The Σ_p and $\tilde{\Sigma}_{error}$ for 3-state phone model are shown in figures 1 and 2 for spectral and temporal domains respectively. In spectral domain, discriminant information ($trace(\Sigma_{error}^{-1} * \Sigma_p)$) is highest in the center state. This is expected as the center state represents the most stable part of the phone. In temporal domain covariances for states 1 and 3 are asymmetric with respect to $t = 0$ due to the fact that these state are shifted in time with respect to the center state. Using these covariances, we derived the spectral and temporal discriminants (refer to section 2). Spectral discriminants using multi-state phone model were similar to those derived from 1-state phone model[2]. They also performed similar to discriminants derived from 1-state phone model. Temporal discriminants using multi-state phone model, however, were different from those those derived from 1-state phone model. They also performed better than discriminants from 1-state phone model. In both domains, discriminants derived from 5-state phone model were similar to discriminants derived from 3-state phone model in the structure and performance.

We further investigated sequential optimization of time-frequency domains using LDA (refer to section 3). The recognition results using these discriminants (sLDA+tLDA) were better than baseline features (WER=6.4%), and discriminants obtained independently from spectral (sLDA) and temporal (tLDA) domains. The performance was further improved to 5.5% using discriminants derived from 3-state phone model. These improvements are significant at $\alpha = 0.05$.

5 Conclusions

In this work, we showed that results from ANOVA can be used to derive discriminant features for ASR. This was done by first grouping sources into useful and harmful sources. Then LDA was used to obtain discriminant features. In our previous work on ANOVA and LDA, the phone was assumed to have unimodal gaussian distribution. The phone model used for recognition, however, is more complex. In this paper, we improved the estimate of phone information using multi-state phone model in ANOVA. The discriminant features using this extension were found to significantly outperform both conventional features and discriminant features obtained from the 1-state phone model.

[2] Interested reader can refer to [5] for a more detail discussion

References

1. S. van Vuuren and H. Hermansky: Data-driven design of RASTA-like filters. Proc. of EUROSPEECH, Greece (1997) 409-412
2. Sachin S. Kajarekar, N. Malayath and H. Hermansky: Analysis of Sources of Variability in Speech. Proc. of EUROSPEECH, Budapest (1999) 343-346
3. Sachin S. Kajarekar, N. Malayath and H. Hermansky: Analysis of Speaker and Channel Variability in Speech. Proc. of ASRU, Colorado (1999)
4. R. Cole and M. Noel and T. Lander: Telephone speech corpus development at CSLU. Proc. ICSLP, (1994)
5. H. Hermansky and N.Malayath: Spectral basis functions from discriminant analysis Proc. of ICSLP, Sydney, (1998)
6. K. Fukunaga: Statistical Pattern Recognition, 2nd ed., Academic Press, San Diego (1998)
7. Thomas M. Cover and Joy A. Thomas: Elements of Information Theory, John Wiley & Sons, Inc (1991)
8. Robert V. Hogg and Elliot A. Tannis: Statistical Analysis and Inference, 5th ed., PRANTICE HALL (1997)

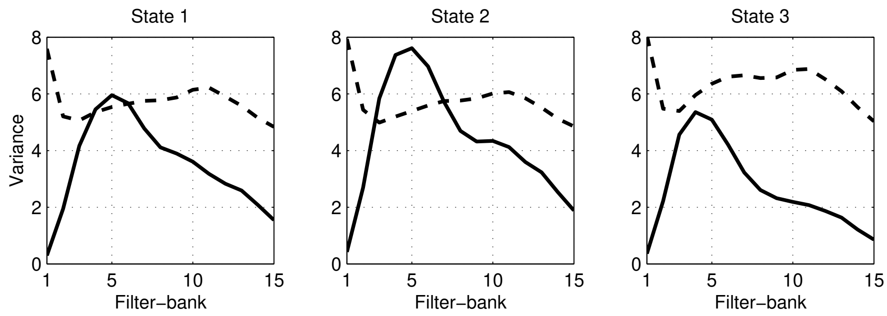

Fig. 1. Spectral Domain: The diagonal of across-class covariance (denoted by solid line) and with-class covariance (denoted by dotted-line) for 3 states. Note the higher phone information in the center state.

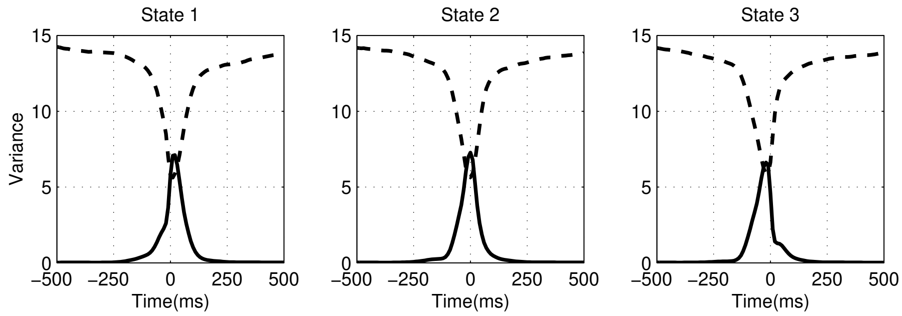

Fig. 2. Temporal Domain: The diagonal of across-class covariance (denoted by solid line) and with-class covariance (denoted by dotted-line) for 3 states. Note the anti-symmetry in the covariances for states 1 and 3.

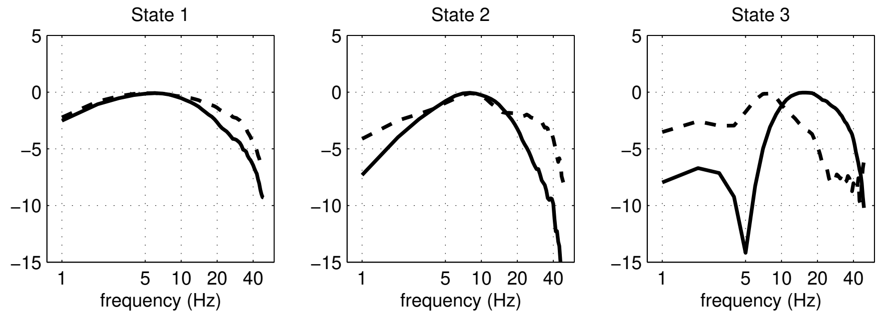

Fig. 3. Frequency response of temporal discriminants: Discriminants using 1-state phone model (demoted by dotted-line) and discriminants using 3-state phone model (denoted by solid line).