6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Analysis of Sources of Variability in Speech

Sachin Kajarekar[1], **Narendranath Malayath**[1] and **Hynek Hermansky**[1],[2]
[1]**Oregon Graduate Institute of Science and Technology**, Portland, Oregon, USA.
[2]**International Computer Science Institute**, Berkeley, California, USA.
email: {sachin,naren,hynek}@ece.ogi.edu

10.21437/Eurospeech.1999-89

Abstract

The variability in the speech signal can be attributed to the following sources: (a) Phonetic content, (b) Speaker and Channel, and (c) Coarticulation or context. In this paper, the variability in speech is decomposed using Two Factor Analysis of Variance (ANOVA) with the above mentioned sources as factors. The speech variability is decomposed in temporal and spectral domain separately and structure of these sources of variability in time-frequency plane is described. Although these factors are not indepdendent, it is shown that they can be studied independently after modeling the interaction between the factors.

1 Introduction

Speech signal contains various sources of information, e.g., phoneme, context, speaker, environment, etc. Hence, the information in speech can be decomposed into the information in these sources. Usually, not all the sources are relevant to the given task and we believe that decomposing the information in speech can help in suppressing the influence of unwanted sources of information. As the information in speech resides in speech variability, decomposing the information implies decomposing the variability.

The variability in speech can be decomposed using different analysis techniques and different choice of factors. In the previous study[1], TIMIT [2] database was analyzed using hierarchically structured Analysis of Variance (ANOVA) technique. This study concluded that the inter-phoneme variability is only 34.1% of the total variability. The remaining within-phoneme variability was attributed to context variability (27.9%), variability due to the position of the frame with a phoneme (26.5%), and the speaker variability (11.6%).

In this paper, Two Factor ANOVA [3] was used. Speaker+channel [1] (referred as SPCH henceforth) and phoneme unit were chosen as the two factors. Since these factors might not be independent, their interaction was studied too. Further, it is shown in section 2 that the "error" term in the Two Factor Analysis models the effect of context variability. Thus, the total variability in speech is decomposed into variability due to the 3 sources of information, 1) phoneme, 2) context, and 3) SPCH.

The paper is organized as follows: Two Factor ANOVA is described in section 2 and the experimental setup is described in section 3. This is followed by the results of the two factor analysis in spectral and temporal domain in Section 4. The analysis was extended to study only the speaker variability in Section 5. Finally, we conclude with the discussion in section 6

2 Two Factor ANOVA

Two Factor analysis is an extension of ANOVA which can be also considered as a one factor analysis. The underline model for the Two Factor Analysis is

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (1)$$

i.e., the mean for a speaker i, speaking a phoneme j, μ_{ij} can be expressed in terms of global mean μ, the mean of the speaker α_i, the mean of the phoneme β_j and the phoneme specific component of speaker γ_{ij}. Using this model and assuming equal number of data points for all speakers and all phonemes, the total sum of squares can be decomposed as follows

$$\begin{aligned} SS(TO) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{...})^2 \\ &= bc\sum_{i=1}^{a}(\bar{X}_{i..} - \bar{X}_{...})^2 + ac\sum_{j=1}^{b}(\bar{X}_{.j.} - \bar{X}_{...})^2 \\ &+ c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2 \end{aligned}$$

[1] Choice of speaker and channel as one factor is a related of the choice of the database. For a database with speaker and channel as labeled sources of variability, it would be possible to do an analysis of their independent effects and their interaction.

$$
\begin{aligned}
&+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{ij.})^2 && (2)\\
&= SS(A) + SS(B) + SS(AB) + SS(E) && (3)
\end{aligned}
$$

where,

X = feature vector
a = no. of speakers
b = no. of phonemes
c = no. of samples for speaker i and phoneme j
SS(TO) = total sum of squares (SS)
SS(A) = SS of factor 1 (phoneme)
SS(B) = SS of factor 2 (SPCH)
SS(AB) = SS of the interaction of factor 1 and 2
SS(E) = SS of error (context and coarticulation)
$\bar{X}_{...} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$ = global mean
$\bar{X}_{i..} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$ = mean of phoneme i
$\bar{X}_{.j.} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c} X_{ijk}$ = mean of speaker j
$\bar{X}_{ij.} = \frac{1}{c}\sum_{k=1}^{c} X_{ijk}$ = mean of phoneme i and speaker j

For our analysis, the above method is modified to take care of unequal sample size ($c \rightarrow c_{ij}$) as follows

$$
\begin{aligned}
SS(A) &= \sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}(\bar{X}_{i..} - \bar{X}_{...})^2\\
SS(B) &= \sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}(\bar{X}_{.j.} - \bar{X}_{...})^2\\
SS(TR) &= \sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}(\bar{X}_{ij.} - \bar{X}_{...})^2\\
&= SS(A) + SS(B) + SS(AB)\\
SS(E) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c_{ij}}(X_{ijk} - \bar{X}_{ij.})^2
\end{aligned}
$$

In words, the variance of phoneme means is modeled as the effect of first factor (SS(B)). Similarly, the variance of SPCH means is modeled as the effect of second factor (SS(A)). The variability within each speaker, speaking a phoneme is the effect of context and coarticulation and it's average is modeled in an error term (SS(E)). The variance of mean of speaker i, speaking phoneme j (SS(TR)), is used to calculate the interaction between the two factors (SS(AB)).

3 Experimental Setup

3.1 Database

OGI Stories [4] database was analyzed with the Two Factor ANOVA. OGI Stories contains about 3 hours of conversational speech which was phonetically hand-labeled. That represents 210 speakers, speaking for about 50 sec each, though different telephone channels. Since, the telephone channels are not labeled, the effect of SPCH is combined in one factor and phoneme class represents another factor in Two Factor Analysis. (Note that the speaker variability is analyzed separately in section 5.) A set of most frequently occurring 38 phonemes were analyzed in this study.

3.2 Features

15 critical band spectrum, calculated using 25 ms window at 100 Hz, was used as the base feature in the spectral analysis.

For the temporal analysis, each band was analyzed independently. A 101 point feature vector was used for the analysis of each band. The phoneme class was modeled using all such 101 point feature vectors which are labeled by the phoneme in its center and span 50 frames on either side [5].

4 Results

4.1 Spectral domain

Figure 1-(a) shows the decomposition of the total variance in spectral domain. It shows that the inter-phoneme variance (Curve II, Figure 1-a) is only 36.2% [2] of the total variance. This indicates that under realistic environments where many speakers speak over varying channels the phonetic variability is significantly lower than the variability caused by the combined effect by SPCH and context. Further, SPCH variability (Curve III, Figure 1-a) is 40.3% and context variability (Curve IV, Figure 1-a) is 23.5% of the total variability.

The variance of the interaction (Curve V, Figure 1-a) is the SPCH information in phoneme after cepstral mean subtraction (CMS). Note that this variance is lower than the context variance. The variance due to the interaction and the context is relatively constant throughout the frequency. Therefore, the peak around 500 Hz suggests the dominance of this frequency band for discriminating between phonemes. This observation is consistent with our earlier work on spectral basis functions [6].

4.2 Temporal domain

The Figure 1-(b) shows the decomposition of total variability for band 5. The decomposition for the other bands have similar trends.

Following observations are made about the structure of the variability:

[2] $\%contribution = trace(factor\_cov)/trace(total\_cov) * 100\%$

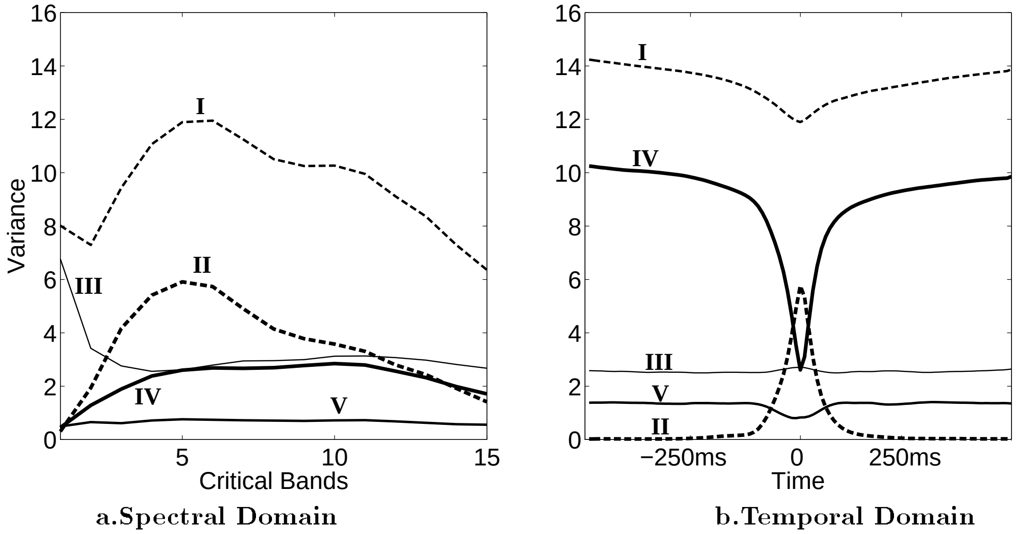


Figure 1: OGI Stories Database: (I) Total variability, (II) phoneme variability, (III) SPCH variability , (IV) context variability, (V) Interaction (between phoneme and SPCH factors) variability

1. the SPCH variability (Curve III, Figure 1-b) is almost constant across time,
2. the effect of phoneme [3] lasts for about 125 ms on either side,
3. the variability of the interaction between phonemes and SPCH (Curve V, Figure 1-b) is the SPCH information after CMS and contributes the least to the total variability(9.0%),
4. context (Curve IV, Figure 1-b) is the most dominating factor and (except in the phoneme) it contributes the most to the total variability (67.2%), and
5. Since only 38 phonemes can occur in the center, but all other phonemes can occur in the context (including silence), there is a dip in the center of the total variability (Curve I, Figure 1(b)).

Observation (1) implies that long term mean removal techniques (e.g., cepstral mean subtraction(CMS)) will suppress the first order effect of SPCH variability.

The interaction between SPCH and context is not modeled explicitly in this analysis. But, curve V in Figure 1-b shows the influence of context on the SPCH information. The variability due to phoneme and SPCH interaction is least at the center. Hence, the center represents the most accurate estimate of SPCH. We believe that an increase in the variability of this interaction away from the center is due to (part of) the context variability getting modeled as the speaker variability.

5 Analysis of Speaker Variability

Since the telephone channel is not labeled in the OGI Stories database, we have done a Two Factor ANOVA of TIMIT database to analyze the speaker variability alone. In order to minimize the effect of context variability on speaker variability, 60 sets of 7 speakers, speaking the same 7 sentences, were analyzed independently. Then, the results of each analysis were averaged. The temporal feature vector and the number of phonemes used for this analysis are as described in section 3.2.

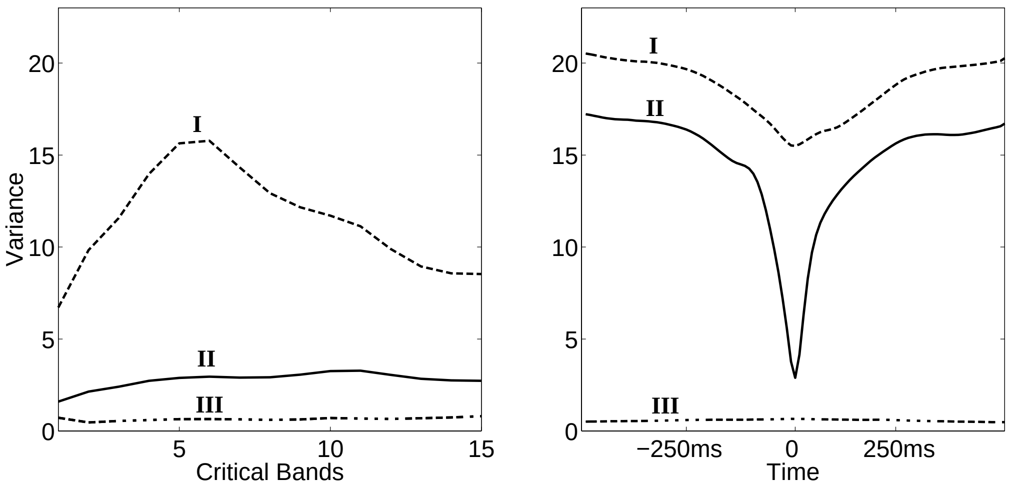
Figure 2 (a) shows the decomposition in spectral domain. Comparing to the SPCH variability in OGI Stories database, the speaker variability (5.6%) in TIMIT is almost constant across frequency. Hence, the dominance of SPCH variability in Figure 1 a(III) at low frequencies can be attributed to the channel variability.

Figure 2 (b) shows the decomposition in temporal domain. The context variability is 78.2% of the total variability. This can be attributed to the design of TIMIT database. It also indicates that the speaker variability is only about 3.0% of the total variability !

6 Summary and Discussion

We have analyzed speech variability using 3 broad factors, 1) phoneme, 2) context, and 3) speaker+channel. The advantage of using these factors is that the results of the analysis can be interpreted in terms of the sources of information in speech. The factors are not independent and interact with each other. We have shown that by modeling the interaction using Two Factor analysis,

[3] average length about 75 ms



a.Spectral Domain **b.Temporal Domain**

Figure 2: TIMIT database. (I) Total Variability, (II) Context Variability and (III) speaker Variability

the factors can be studied more accurately.

Comparing our results of the decomposition in spectral and temporal domain with [1], few comparisons can be made: 1) Our results on the contribution of phoneme class variability in spectral domain agree with [1]. They also indicate that the effect of the phoneme variability is spread beyond the length of the phoneme. 2) Similarly, the contribution of the context variability in the spectral domain is in consent with the similar conclusion in [1]. 3) There is a difference in the contribution of speaker variability of our study and [1] and it can be attributed to a possible influence of context variability on speaker variability in [1]. 4) Finally, we have noted similar results of the contribution for the context variability in time domain as the variance of the phoneme sub-segment in [1].

The comparison of the speaker variability (3%) from TIMIT database and the SPCH (9%) variability from OGI Stories database indicates that the utterance mean contains more information about the channel than speaker.

The fact that different sources of variability have different structure in frequency domain, spectral basis functions can be designed to suppress some sources of variability [6]. Similarly, the different structure of the sources of variability in temporal domain allows for the design of temporal RASTA filters to suppress the effect of some sources of variability [5]. However, the fact that there is an interaction between the sources, the scope of the linear discriminant techniques is limited.

References

[1] Don X. Sun and Li Deng, "Analysis of acoustic-phonetic variations in fluent speech using timit," in *Proc. of ICASSP*, Detroit, 1995, pp. 201–204.

[2] R. H. Kasel L. F. Lamel and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. of the DARPA Speech Recognition Workshop*, 1987, pp. 26–32.

[3] Robert V. Hogg and Elliot A. Tannis, *Statistical Analysis and Inference*, PRANTICE HALL, fifth edition, 1997.

[4] T. Lander R. A. Cole, M. Noel, "Telephone speech corpus development at cslu," in *Proc. of ICSLP*, 1994.

[5] S. van Vuuren and H. Hermansky, "Data-driven design of rasta-like filters," in *Proc. of EUROSPEECH*, Greece, 1997, pp. 409–412.

[6] Hynek Hermansky and Narendranath Malayath, "Spectral basis functions from discriminant analysis," in *Proc. of ICSLP*, Sydney, 1998.

[7] Carlos Avendano, *Temporal Processing of Speech in a Time-Feature Space*, PhD Thesis, Oregon Graduate Institute of Science and Technology, 1997.

[8] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for speech intelligibility in auditoria," *JASA*, vol. 77, pp. 1069–1077, 1985.