

ANOVA IN MODULATION SPECTRAL DOMAIN

Sachin S. Kajarekar¹, Narendranath Malayath¹, and Hynek Hermansky^{1,2}

¹ Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA.

² International Computer Science Institute, Berkeley, California, USA.

ABSTRACT

The sources of variability in modulation spectral domain are studied using ANOVA. It is shown that the speaker+channel variability and phoneme variability are separable in modulation spectral domain. Further, the results of ANOVA are used to design the temporal filters for the speech and speaker recognition task. The observations about the nature of the variability sources were conformed by the frequency response of these temporal filters.

1. INTRODUCTION

It is the variability in speech that carries the information. This variability has several sources, the main ones being the linguistic message, the speaker, and the communication channel used for speech capture, and/or its storage and/or transmission. The linguistic message produces 1) variability due to different phoneme values and 2) the variability within a phoneme due to different phonemic context.

The variability due to the different sources is calculated as using analysis of variance (ANOVA) [1] as follows. First, all the feature frames¹ in the database, labeled by the same phoneme, are collected and their mean is calculated. This is termed as the mean of the phoneme. A similar procedure is used to calculate the mean of the different phonemes, speakers and channels. The variance of phoneme, speaker and channel means is termed as phoneme, speaker and channel variability respectively. Finally, for a given speaker, the variance within a phoneme is the effect of phoneme context. The average of this variance over all phonemes and speakers is calculated and is termed as the context variability.

Using ANOVA we have studied the variability in spectral and temporal domains [2, 3]. The main observations from the spectral studies were that the maximum of phoneme variability is at frequencies around 500 Hz while the speaker and channel variability is distributed quite evenly throughout all frequencies. The

temporal studies indicated a spread of the phoneme-related variability within about 250 msec around the target phoneme frame while the speaker and the channel variability is almost constant across time. The phoneme-specific ANOVA indicated that the relative contributions of different variability sources within individual phonemes are different and sonorants exhibit higher speaker variability than obstruents.

Although the analysis in temporal domain was used to design the filters in modulation spectrum², the structure of the variability sources in the modulation spectral domain could not be derived from temporal ANOVA. Instead, it was hypothesized from the frequency response of the temporal filter.

In this paper, we have used a series of bandstop filters to analyze the variability sources in the modulation spectrum domain. Better understanding of the variability in this domain may allow for a design of RASTA filters which could alleviate and/or enhance contributions of various variability sources.

The paper is organized as follows: section 2 describes the experimental setup for the ANOVA in modulation spectrum and the conclusions from the experiment. We describe the temporal filter design for speech and speaker recognition and compare the filter characteristics with the results of ANOVA in section 3. Finally, we present conclusions in section 4.

2. ANOVA IN MODULATION SPECTRAL DOMAIN

2.1. Experimental Setup

We have used OGI Stories database [4] for ANOVA. The database contains about 3 hours of conversational speech which is labeled by OGIbet. The database has 210 speakers, speaking for about 50 sec each, through different telephone channels. Since the conversations were not labeled by the telephone channel, the effect of speaker and channel is combined into a single source.

¹section 2 describes the features used for the analysis

²The modulation spectrum is the spectrum of time trajectory of spectral envelope.

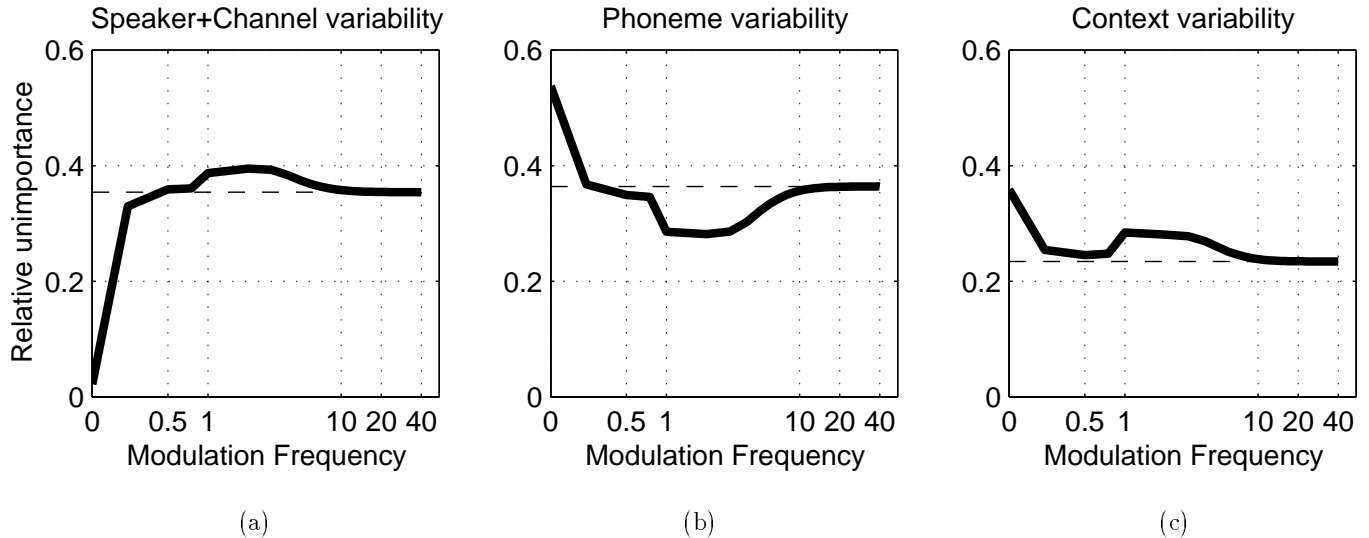


Figure 1: Decomposition of variability in modulation spectral domain.

Band-stop filters of length 1001 points (10 sec) were designed for the experiment. Each band-stop filter has a center frequency f_i such that $f_i = 0, 0.25, 0.5, 1, 2, \dots, 40$ Hz and bandwidth of 1Hz if $f_i \geq 1$ Hz and 0.5Hz if $f_i < 1$ Hz. The 15 log critical band energy spectrum, calculated using 25 ms window at 100 Hz, was used as the base feature. The time trajectories of all the bands were filtered by one filter with center frequency f_i . Finally, analysis of variance [2] was performed on the filtered features in spectral domain. The total (S_{total}) variance was defined in terms of phoneme ($S_{phoneme}$), speaker+channel (S_{sp+ch}), and context variance ($S_{context}$) as

$$S_{total} = S_{phoneme} + S_{sp+ch} + S_{context}$$

2.2. Discussion of results

The results of the bandstop filtering in modulation spectral domain are shown in the Fig.1. The x-axis represents the modulation frequency component that was suppressed using a band-stop filter and y-axis represents the source variance. The source variance is normalized using the total variance (S_{total}) to make the results invariant to the gain of the filter. The index "0" on x-axis represents the source variance after utterance based mean subtraction(MS). The variability at the frequency f_i is the relative residual variability after removing the modulation frequency component f_i . The dotted-line in the figures represents the variability without any filtering. If the variability drops below the dotted line at f_i then we conclude that f_i carries the variability due to the source.

The figure shows that the modulation frequency components around 2 Hz carry the most phoneme variability (Fig.1(a)). This result is validated by the characteristic of temporal filter designed in the next section. The speaker+channel variability lies between 0-0.5 Hz (Fig.1(b)) and is not affected by filtering out the modulation frequency components higher than 1 Hz. This observation is consistent with our earlier observation [3]. Context variability never goes below the variability without filtering. (Fig.1(c)). This means that the modulation frequency domain is not the right domain to characterize the context variability. The figures also indicate that the band-stop filtering beyond 20 Hz does not affect the sources of variability. Thus, there is almost no variability in modulation spectrum beyond 20 Hz. This observation is consistent with the recent down-sampling experiments where it was empirically shown that the speech and speaker recognition performance is not affected by down-sampling the feature stream by a factor of 3-4 [5, 6].

3. TEMPORAL FILTERS FOR SPEECH AND SPEAKER RECOGNITION

3.1. Design of temporal filters

The ANOVA has shown that the variability about phoneme is present beyond the phoneme duration. It is also shown that the mean of a long utterance contains channel variability as well as speaker variability. In this section, we show that the results of ANOVA can be used to design data-driven temporal filters for speech and speaker recognition.

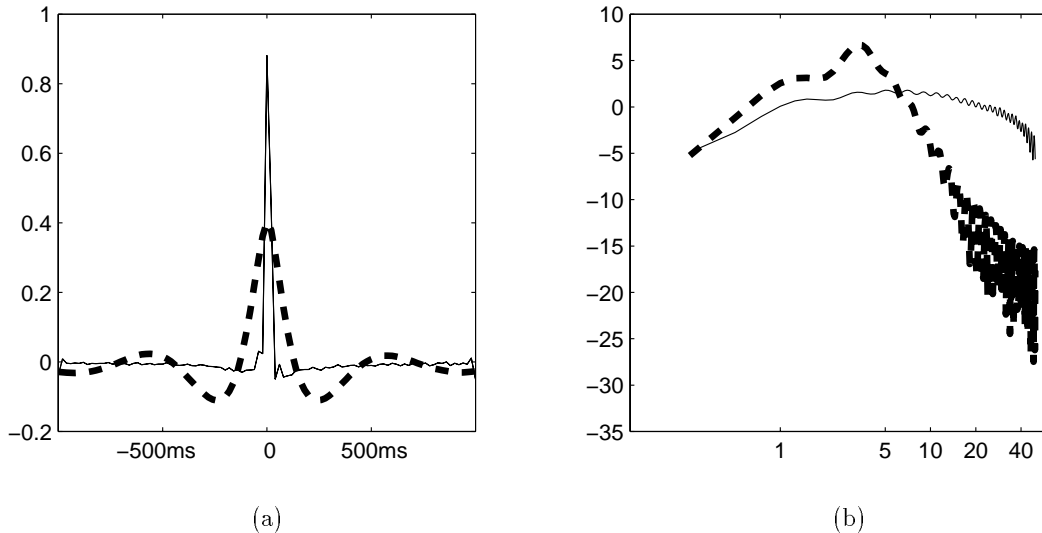


Figure 2: Impulse response and frequency response of Temporal filters designed for speech recognition (solid line) and speaker recognition (dash line)

The design of temporal filters is described in detail in [7, 8]. The general design method can be described as follows. First, the variability in speech is decomposed into the phoneme, context, and speaker+channel variability using ANOVA. These variability sources are grouped into the "useful" variability and "harmful" variability for a given task. The "useful" and "harmful" variability is used by Linear discriminant analysis (LDA) [7, 8] to generate the temporal filters (referred as linear discriminants). The projections of the temporal trajectories of the base features on these linear discriminants are used as the features for the given task.

For speech recognition, the useful source of variability is phoneme variability and the context and speaker+channel variability is the harmful variability. For speaker recognition, the phoneme variability is the useful variability as the useful speaker variability is defined with respect to phoneme variability. The channel variability becomes the harmful variability for the speaker recognition task. The most important linear discriminant for the task of speech and the speaker recognition is shown in Fig.2

3.2. Filter characteristics and ANOVA

The characteristics of linear discriminants can be understood by studying the filter that preserves phoneme variability under no constraint and the filter that preserves the phoneme variability while suppressing context variability (Fig.3). Consider the filter which is derived only from the phoneme variability (Fig.3 dot-

ted line). The concentration of the phoneme variability in 2-3 Hz conforms our observation about nature of phoneme variability. In conclusion, evolution of phoneme in time is a slow process at the rate of about 2-3 Hz.

If the constraint of speaker+channel variability as "harmful" variability is added (Fig.2 dotted line) then the filter suppresses the variability below 1 Hz. Similar filter has been shown to outperform the conventional techniques like MS and RASTA filtering in speaker recognition task [8]. This conforms our earlier observation about the modulation frequencies which carry the speaker+channel variability.

If the constraint of context variability as "harmful" variability is added (Fig.3 solid line) then the filter reduces its temporal width which increases the bandwidth of the filter. Finally, Under the constraint of speaker+channel and context variability as "harmful" variability, the resulting filter is a bandpass filter (Fig.2 solid line). The high pass nature of the filter suppresses speaker+channel variability and the low pass nature suppresses the context variability. This filter have been shown to outperform conventional filtering techniques like MS and RASTA filtering in speech recognition [9].

4. CONCLUSION

It is shown that the speaker+channel and phoneme variability are separable in modulation spectral domain. The speaker+channel is dominant below 1Hz and phoneme variability is dominant in 2-3 Hz. These

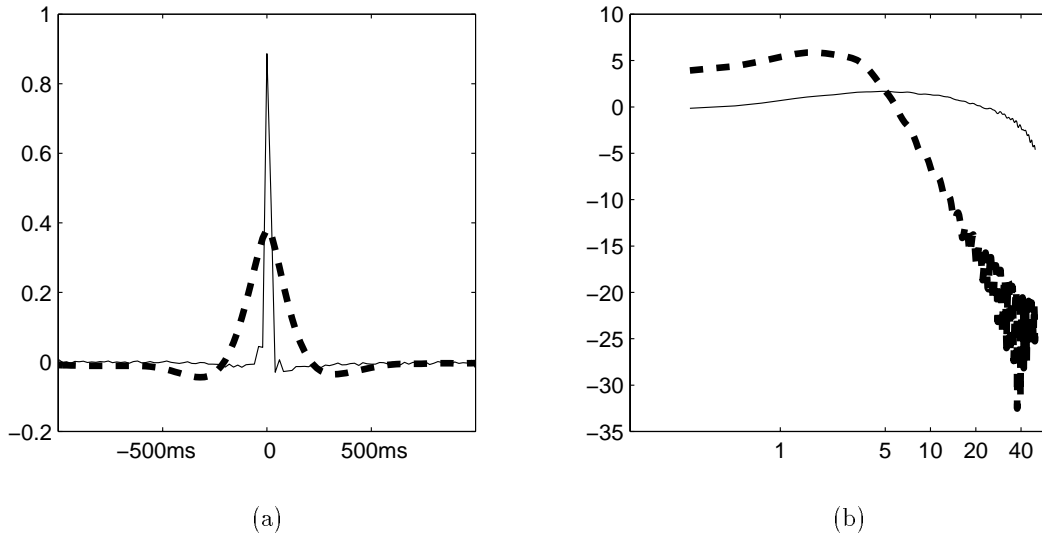


Figure 3: Impulse response and frequency response of Temporal filters preserving phoneme variability(solid line) and phoneme variability under the constraint of suppressing context variability(dash line)

variability sources can be used to design the filters for the task of speech and speaker recognition. The characteristics of these filters conform our observations derived from ANOVA.

Since OGI stories databases is not labeled by the telephone channel, the speaker+channel variability could not be segregated further. It is our hypothesis that the variability at 0 Hz in modulation spectrum is channel variability. The variability greater than 0 Hz and less than 1 Hz is the speaker variability. More analysis needs to be done to verify this hypothesis. The context variability in modulation spectral domain and spectral domain does not have a structure. So, we cannot describe the nature of context variability. Further analysis needs to be done to understand the nature of context variability.

The limitations of the ANOVA are 1) it is assumed that speech variability can be decomposed into the variability due to sources and 2) the sources were modeled using unimodal gaussian distribution. Both of these limitations are addressed in [10]. It computes the mutual information between phoneme labels and the features without any assumption about feature distribution. It is encouraging to find very similar structure of source variability from mutual information as obtained from the ANOVA.

REFERENCES

- [1] Robert V. Hogg and Elliot A. Tannis, *Statistical Analysis and Inference*, PRANTICE HALL, fifth edition, 1997.
- [2] Sachin Kajarekar, Naren Malayath and Hynek Hermansky, "Analysis of sources of variability in speech," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 343-346.
- [3] Sachin Kajarekar, Naren Malayath and Hynek Hermansky, "Analysis of speaker and channel variability in speech," in *Proc. of ASRU*, Colorado, 1999.
- [4] T. Lander R. A. Cole, M. Noel, "Telephone speech corpus development at csu," in *Proc. of ICSLP*, 1994.
- [5] Sarel V. Vuuren, *Speaker Verification in Time-Feature Space*, PhD Thesis, Oregon Graduate Institute of Science and Technology, 1999.
- [6] Hynek Hermansky and Pratibha Jain, "On Down-sampling speech representation in ASR," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 343-346.
- [7] Sarel van Vuuren and Hynek Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *ICSLP*, Sydney, Australia, 1998.

- [8] N. Malayath, S. Kajarekar and H. Hermansky, "Speaker verification with channel normalizing filters," in *Proc. of NIST Speaker Recognition Workshop*, University of Maryland, Washington, 1999.
- [9] S. van Vuuren and H. Hermansky, "Data-driven design of rasta-like filters," in *Proc. of EUROSPEECH*, Greece, 1997, pp. 409–412.
- [10] H. Yang, S. van Vuuren and H. Hermansky, "Relevance of timefrequency features for phonetic classification measured by mutual information," in *Proc. of ICASSP*, Phoenix, Arizona, 1999.