ANALYSIS OF ACOUSTIC-PHONETIC VARIATIONS IN FLUENT SPEECH USING TIMIT

Don X. Sun [1] *and Li Deng* [2]

[1]Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794.
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

ABSTRACT

In this paper, we propose a hierarchically structured Analysis of Variance (ANOVA) method to analyze, in a quantitative manner, the contributions of various identifiable factors to the overall acoustic variability exhibited in fluent speech data of TIMIT processed in the form of Mel-Frequency Cepstral Coefficients. The results of the analysis show that the greatest acoustic variability in TIMIT data is explained by the difference among distinct phonetic labels in TIMIT, followed by the phonetic context difference given a fixed phonetic label. The variability among sequential sub-segments within each TIMIT-defined phonetic segment is found to be significantly greater than the gender, dialect region, and speaker factors.

1. INTRODUCTION

It has been known from many years of speech research, both theoretical and empirical, that speech variability at the acoustic level is overwhelming. In fact, such variability constitutes the major obstacle to the construction of machines for high-performance speech recognition even with the state-of-the-art technology.

In addition to the demonstrations of the overwhelming amount of speech variability, the various factors that contribute to such variability have also been identified and studied by speech scientists and engineers. However, the quantitative aspects of these factors have been conspicuously lacking in the literature, and, to our best knowledge, all previous studies on speech variability have been limited either to small subsets of speech data or to only the qualitative aspect of the study in a descriptive fashion. Although the acoustic-phonetic speech database TIMIT has been around for sometime and is ideally suited to the study of the quantitative aspect of the variability for all classes of speech sounds, no such a study has been undertaken in the past.

The purpose of this paper is to report the results of our recently conducted comprehensive study on the acoustic variability of fluent speech and on the related factors using TIMIT. In addition to the conclusions from this study and their implications for automatic speech recognition system design, one key contribution of our study is a novel hierarchically structured Analysis of Variance (ANOVA) method developed from this study that enables the conventional ANOVA analysis to be conducted in an efficient and meaningful way. Using the hierarchically structured ANOVA method, we have been successful in decomposing the global acoustic variability in fluent speech according to hierarchically organized, phonetically and speaker related factors defined in the TIMIT database.

2. HIERARCHICALLY STRUCTURED ANALYSIS OF VARIANCE METHOD

Analysis of Variance (ANOVA) ([7]) is an effective statistical method in analyzing quantitative responses from experimental units. The main idea of ANOVA is to decompose the overall variation among the observations into various components corresponding to the factors involved in the experiments. The purpose of such an analysis is to assess the relative significance of the various factors in affecting the response. There are many different types of ANOVA models depending on the structure of the experimental factors. The model with two nested factors is illustrated here as the building block of our proposed method. Consider the following model with two factors A and B,

$$Y_{ij} = \mu + \mu_i + \mu_{j|i} + \epsilon_{ij}, \quad i = 1, \cdots, I; \;\; j = 1, \cdots, J(i);$$

where Y_{ij} is the observed response at ith level of factor A and jth level of factor B and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the corresponding experimental error. The parameter μ_i represents the effect of ith level of factor A and $\mu_{j|i}$ represents the effect of jth level of factor B given that factor A is at level i. For instance, we can consider gender of the speakers as factor A and individual speakers as factor B which is nested with the levels of factor A. The total variation in the observed responses can be decomposed into two sum-of-square terms corresponding to the two factors:

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{I}\sum_{j=1}^{J(i)} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J(i)} (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i=1}^{I} J(i)(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= SS_A + SS_B \end{aligned}$$

The ratio between SS_A and SS_B indicates the relative significance of the two factors in affecting the response.

For studying the various aspects of the acoustic variabilities in fluent speech, ANOVA is a very suitable method. The experiment for our analysis is undertaken using

Figure 1. Factors at phonetic level

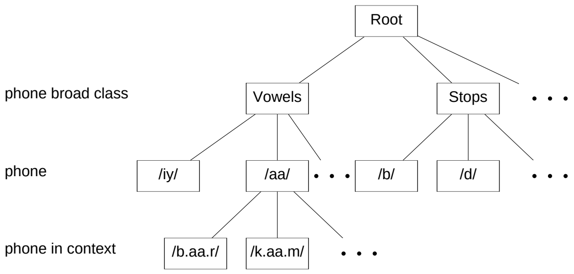
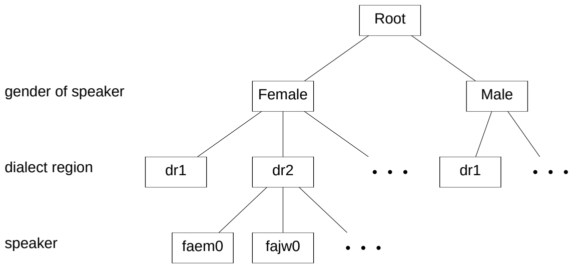


Figure 2. Factors at speaker level (I)



the carefully designed acoustic-phonetic speech database TIMIT. TIMIT is ideally suited for such an experiment since it involves balanced selection of a full range of variability factors including phonetic contexts, dialect regions of the speakers, gender of the speakers, etc.

The factors representing various aspects of the variability in fluent speech can be organized in the hierarchical structures as illustrated in Fig. 1–4.

In Fig. 1, the variation in the root node represents the overall variation in the observations. It can be decomposed into within-class variation and between-class variation. Within each phonetic broad class, the variation can be further decomposed into within-phoneme and between-phoneme variations. Similarly, the variation within each phoneme can be decomposed into within- and between-phonetic-context variations. In Fig. 2 and 3, we give two possible arrangements for the factors of gender, dialect region, and speaker, since there is no strict nesting relation between the factors of gender and dialect region. Finally, the model structure shown in Fig. 4 is used to study speech variability at three levels with the following descending order: token level, sub-segment level, and frame level. A token is a sequence of observations that correspond to a specific speech units such as a phoneme. A sub-segment is a portion of a token; temporal variation of sub-segments that constitute a token shows dynamic properties of the speech signal. A frame is one single observation (vector), typically covering 10 ms of the speech signal, obtained from a signal preprocessor. For simplicity in our analysis, we artificially divide each token into three uniformly spaced sub-segments: initial, middle, and final sub-segments.

Figure 3. Factors at speaker level (II)

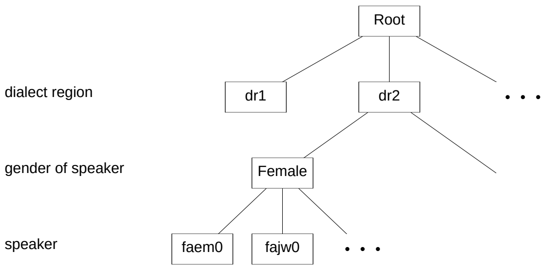
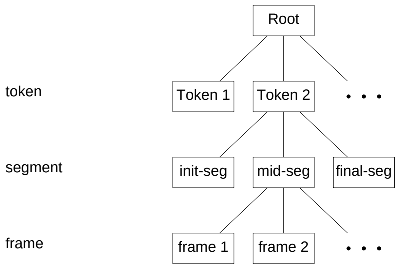


Figure 4. Factors at frame level



The tree structures in Fig. 1–4 can be combined to form richer models for the ANOVA analysis. On the other hand, the layers in these tree structures can also be omitted to reduce the size of the analysis models.

3. AN EFFICIENT ALGORITHM FOR HIERARCHICAL ANOVA USING SEQUENTIAL DATA

In principle, the ANOVA can be obtained by a regression analysis with proper coding of the factor levels using orthogonal contrasts ([7, 1]). However, when the number of factors and number of levels of the factors become large, it is computationally prohibitive to carry out the ANOVA in the traditional way. In this paper, we propose a new approach to the computation of ANOVA without requiring large amount of computer memory. By taking advantage of the nested hierarchical structure of the ANOVA model, we can isolate different sources of variations and devise an efficient updating procedure for computing the sum-of-square terms corresponding to the layers in the tree structure.

For each incoming observation, a path in the tree structure is identified or a new path is created if the factor levels corresponding to the observation have never occurred before. The observation vector is accumulated into each node on the path and the counter of number of observations of that node is incremented. When all the observations are updated, the sum-of-square terms of each node can be updated by the statistics of all the child node in a recursive fashion. In fact, since TIMIT database is organized by speakers, we can update the sum-of-square statistics corresponding to all the layers below the "speaker" factor when the speech data of one speaker has been completely acquired. Then, all the branches below the current "speaker" node can be safely pruned. Therefore, at any time, there is only one active sub-tree structure below the "speaker" level. This implies that the amount of computer memory required for the computation of the entire database is almost equivalent to that for one speaker. This algorithm is very effective for ANOVA analysis of large data sets where the conventional method would fail.

Table 1. Contribution of individual factors

Cl	17.6%
Ph	16.5%
Ct	27.9%
Dr	**9.2%**
G	**1.6%**
Sp	0.8%
Tk	0.6%
Sg	16.5%
Fr	9.4%

Cl	17.6%
Ph	16.5%
Ct	27.9%
G	**4.2%**
Dr	**6.6%**
Sp	0.8%
Tk	0.6%
Sg	16.5%
Fr	9.4%

4. RESULTS AND CONCLUSIONS

In this section, we present the major results and conclusions obtained from this study. The results are expressed in terms of the percentage of contribution of each factor to the overall variation. The raw sum-of-square statistics in the ANOVA are not presented since they are not so important to the objective of this study. The ANOVA are computed based on the observation vectors of the Mel-Frequency Cepstrum Coefficients (MFCC) C_1–C_7. All the MFCC vectors are mean centered and standardized.

To simplify the presentation of the tables, we adopt the following abbreviation for the factors:

Cl	=	phoneme broad class
Ph	=	phoneme unit
Ct	=	phoneme-in-context
G	=	gender of the speaker
Dr	=	dialect region of the speaker
Sp	=	speaker
Tk	=	token of one speech unit
Sg	=	sub-segment with a token
Fr	=	frame with each sub-segment of the token

4.1. Contribution of the individual factors

The most important objective of this study is to assess the effect of various factors in contributing to the overall variation in speech signals. In Table 1, we present the ANOVA results based on two slightly different model structures. The difference between the two models (the factors in bold face) is the nesting order between the factors "gender" and "dialect region".

From this table, we can draw the following conclusions:

1. About (17.6%+16.5%) = 34.1% of the total variation is explained by the differences among the phoneme units. This implies that modeling speech signals at the phoneme level will lose a considerable amount of information.

2. Among the remaining part of the variation, 27.9% can be further explained by the variation among phonemes in different phonetic contexts. This indicates great potential in modeling context dependent speech units for improvement of speech recognition systems. This conclusion is consistent with many research works in context dependent speech modeling ([6, 5, 3]).

3. Below the phonetic level, (9.2%+1.6%+0.8%) = 11.6% out of the remaining 38% is explained by the variation among the speaker. This result shows that although modeling speaker variation for speaker independent speech recognition is usually regarded as a difficult problem, it is less severe than the variation caused by the context dependency of the phonetic units.

4. Finally, there is a surprisingly large variation among different sub-segments within each token. About 16.5% out of the remaining 26.5% variation is due to the sug-segment effects. This suggests that speech signals within a phonetic segment is far from stationary and more research on modeling the dynamic patterns of speech will help to improve the performance of existing speech recognition systems [4, 2].)

Table 2. Effect of context-dependent speech units

	M1	M2	M3	M4
Unit	34.1%	37.2%	49.2%	74.4%
G	3.0%	3.3%	5.8%	5.3%
Dr	8.6%	10.3%	14.9%	8.7%
Sp	7.1%	8.5%	6.3%	1.2%
Tk	21.4%	31.3%	14.5%	1.0%
Sg	16.5%	6.6%	6.6%	6.6%
Fr	9.4%	2.8%	2.8%	2.8%

4.2. Effect of context-dependent speech units

As we can see from the results in the previous section, phonetic context variation is a major difficulty in speech modeling. In this section, we use the ANOVA method to assess the capability of various context-dependent speech units in explaining the total variation in speech.

We consider the following four models: two context independent and two context dependent models.

M1: phonemes as the basic speech units (61 in total);

M2: three sub-segments within a phoneme segment as the basic speech units (183 in total);

M3: three sub-segments within a TIMIT phone segment modeled by a pair of left- and right-diphones and a center phone units (3149 in total);

M4: three sub-segments of a triphone unit as the basic speech units (15971 in total).

From Table 2, we observe that the speech units in "M1" have a rather low contribution to the overall variation (34.1%). "M2" does not improve this contribution much (from 34.1% only to 37.2%). The context dependent model based on diphone units "M3" takes a significant leap from the context independent model. The variation explained by the diphone units goes from 37.2% to 49.2%. As a step further, the context dependent model based on triphone units "M4" improves even more significantly over the model based on diphone units. The variation explained by the triphone units goes up to 74.4%.

Although "M4" shows the greatest contribution to the overall variation in the fluent speech, the number of speech units is too large (around 16,000 in this example) to be reliably estimated from any moderate amount of training data. Our result suggests that better modeling of context dependent behaviors in speech using more parsimonious models

Table 3. ANOVA for individual phonemes

Ph	Ct	Sp	Tk	Sg	Fr
aa	48.1%	19.6%	0.2%	20.2%	11.8%
ae	52.6%	15.9%	0.1%	19.2%	12.2%
ah	46.8%	23.3%	1.3%	19.2%	9.4%
ao	51.8%	12.9%	0.5%	22.5%	12.3%
b	30.0%	29.4%	4.0%	28.8%	7.9%
d	30.4%	33.6%	4.3%	24.7%	7.0%
em	48.7%	7.7%	0.0%	27.2%	16.4%
en	49.8%	8.5%	0.0%	25.6%	16.1%
⋮					

should benefit greatly the design of speech recognition systems.

4.3. ANOVA for individual phonemes

We have also performed ANOVA for each individual phoneme in terms of context, speaker, token, and segment variations. Due to space limitation, we can only present the results for a fraction of the phonemes in Table 3. The general conclusion from this analysis is that the contextual variation within each individual phoneme dominates all the TIMIT-defined factors we put into our analysis.

5. DISCUSSIONS

It is well known that there are large acoustic variations in spoken language, where one most significant contribution to the variation of speech is the intrinsic difference among distinct underlying linguistic units. This distinction, intermixed with the acoustic variations resulting from different linguistic units, carries linguistically meaningful information for human speech communication.

However, other non-linguistic factors that also contribute significantly to the global speech variability. First, since fluent speech is not a simple concatenation of phonetic symbols, it's inevitable that phonetic units defined at the phoneme level be easily influenced by adjacent units. Second, even for the same phonetic symbol in the same phonetic context, the acoustic signals are generally heavily influenced by a number of speaker-related factors (gender, dialect region, individual vocal tract differences, and speaking style, etc.). These factors also carry certain amount of information (which is useful for speaker identification), but for the objective of speech understanding, these variations will harm the performance of speech recognition systems.

To improve the performance of speech recognition systems, there have been numerous efforts devoted to reducing and separating the undesirable variations in natural speech. Although the importance of quantitative understanding of the numerous sources of these variations has been well recognized, any systematic study of this kind requires sophisticated statistical skills. Although the conventional ANOVA analysis appears appealing to achieve our goal of quantitative analysis, the complexity of the variation factors and the large amount of speech data in TIMIT do not permit a straightforward adoption of the conventional technique. As one main contribution of this study, we have developed a simple-to-implement recursive algorithm, as described in Section 3, that updates the various variation components in ANOVA in a sequential and adaptive manner. This algorithm has not appeared in the statistical literature and is essential for us to obtain the quantitative results described in Section 4.

The current study, together with the preliminary conclusions reached in Section 4, has been limited in several ways. First and foremost, our entire analysis has been based on the assumption that the acoustic data of speech are represented by the mel-frequency cepstral coefficients. Second, the speaking style in TIMIT data is relatively uniform from one speaker to another and from one sentence to another, and is far from that of natural conversational speech. Third, the syntactic and semantic structures of the TIMIT sentences are far from those of natural conversational speech. Despite of these limitations, our results nevertheless serve to provide useful insights to the understanding of the roles of various components of speech recognizers in contributing to the ultimate speech recognition performance. For instance, the surprisingly large amount of variation, found in our ANOVA analysis, across sequentially ordered subsegments of speech within a segment of the TIMIT-defined phonetic label is indicative of the significant role of dynamic modeling in speech recognizer design. (Such a role has been empirically demonstrated in our earlier study [2].)

Acknowledgements

The support for this work is provided by Natural Sciences and Engineering Research Council of Canada and by National Science Foundation Grant Number DMS-9057429.

REFERENCES

[1] J. M. Chambers and T. J. Hastie. *Statistical models in S.* Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[2] L. Deng, M. Aksmanovic, D. X. Sun, and C. F. J. Wu. Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, 1992.

[3] L. Deng and D. X. Sun. A statistical framework for automatic speech recognition using the atomic units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America*, 95(5):2702–2719, 1994.

[4] O. Ghitza and M. Sondhi. Hidden markov models with templates as nonstationary states: an application to speech recognition. *Computer Speech and Language*, 7(2):101–119, 1993.

[5] M. Hwang, X. Huang, and F. Alleva. Predicting unseen triphones with senones.

[6] K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX System.*

[7] H. Scheffe. *The analysis of variance.* Wiley, New York, 1961.