

# ON THE DISTRIBUTION OF INFORMATION AND INTRINSIC VARIABILITY FOR CLASSIFICATION OF COARTICULATED VOWELS

*Sorin Dusan*

Speech and Language Processing Laboratory, Center for Advanced Information Processing  
Rutgers University, Piscataway, NJ, 08854, U.S.A.  
sdusan@caip.rutgers.edu

## ABSTRACT

It is known that the information necessary for the identification of coarticulated vowels is distributed throughout the duration of the vowels and the adjacent phonemes. This information is embedded in the speech signal in various acoustic patterns: static, dynamic, and temporal. A recent study identified seven types of acoustic patterns that might be exploited by listeners in the identification of coarticulated vowels. This paper extends the previous study and focuses on two problems. First, it presents a quantitative analysis of the underlying distribution of the acoustic information throughout vowels and adjacent consonants by employing vowel classification experiments. Second, it presents a quantitative analysis of the within-groups and between-groups sources of variability that make the total intrinsic variability of speech, and shows that the results of such analysis correlate with and predict the results obtained in the vowel classification experiments that reflect the distribution of information. The findings of this paper may be important for automatic speech recognition and suggest some basic improvements to such techniques.

## 1. INTRODUCTION

A well cited study of the distribution of formant frequencies at the center of vowels showed significant overlap between regions corresponding to adjacent vowels [1]. This led to the conclusion that vowels cannot be completely represented by a single spectral section at the target position [1]. A subsequent study suggested that formant direction and rate of change at formant transition also play an important role in vowel recognition [2]. Another study showed that the identification of vowels by listeners is more accurate when vowels are coarticulated with consonants than when they are uttered in isolation [3]. Other studies focused on the identification of vowels by listeners based on dynamic and temporal acoustic cues by employing silent-center or truncated syllables [4], and [5]. A recent study

analyzed the vowel discrimination properties of various spectral (static and dynamic) and temporal cues along the continuous dimension of time for nine American English vowels uttered in three left- and three right-consonant contexts [6]. This study used speech data from the TIMIT database [7] and identified seven types of acoustic patterns (cues) that might be used by listeners for the identification of coarticulated vowels.

A number of studies in the past focused on revealing the distribution of information correlated with the phoneme identity across various acoustic features, e.g. [8], [9], [10], and [11]. All these studies employed measures from information theory to reveal such distributions of information. Such analyses are important because they provide a measure of usefulness or importance of each feature for phonetic classification. However, these studies aimed mainly at ranking all the examined features or combinations of two features and selecting the best features in a long feature vector for classification.

This paper extends the previous work and focuses on two related problems. First, the paper presents a quantitative, indirect analysis of the distribution of the acoustic information correlated with vowel identity along various types of patterns by evaluating the accuracy of an automatic vowel classification method for each type of pattern. Unlike in listening experiments, where it is difficult to restrict the identification of vowels to a single cue at a time, the automatic classification used here enables a detailed analysis for each type of pattern or acoustic feature. This analysis aims at quantifying the information correlated with the vowel identity within and outside the currently accepted boundaries of vowels. Second, the paper presents an analysis of the sources of intrinsic variability that can explain and predict the shape of the underlying distribution of vowel information. This research aims at finding ways of improving current HMM technologies.

## 2. SPEECH DATA

The speech material used in this study is from the TIMIT American English speech corpus [7] and is the same as in the previous study [6]. It consists of all the biphone tokens

from the training part of the corpus (462 speakers) corresponding to nine vowels uttered in three left-consonant (CV) and three right-consonant (VC) contexts. The vowels are /aa/, /ae/, /ah/, /ao/, /eh/, /ih/, /iy/, /uh/, and /uw/, and are selected to match those used in a previous study on dynamic specification of coarticulated vowels [4]. For more variety in the consonant context the three consonants are selected from three different classes (plosive, fricative, and nasal) and are /b/, /sh/, and /m/, respectively. Since TIMIT provides separate segmentations and labels for the stop closure (/bcl/) and stop release (/b/) these separate segments were used in the /b/\_V and V\_/bcl/ biphones, respectively. In total there are 27 CV and 27 VC biphone types. The numbers of tokens in the biphone types vary between 4 and 522 with the average of 102 per biphone type.

### 3. DISTRIBUTION OF INFORMATION: METHOD

The analysis of the distribution of acoustic information correlated with vowel identity is done by evaluating the vowel classification accuracy at various positions along the CV and VC trajectories and for various acoustic patterns. Since the classification of the nine vowels is done individually for each pattern type, the classification accuracy can indirectly serve as an approximate measure of the distribution of the acoustic-phonetic information associated with these patterns. Previous studies analyzing the distribution of information correlated with vowel or phoneme identity employed the mutual information (MI) or the joint mutual information (JMI), e.g. [8], [9], [10], and [11]. They used phonetic classifications mainly to confirm that the estimated distributions reflect the classification accuracies. The use of mutual information to reveal such distribution is theoretically motivated by the information theory. However, for tasks involving multivariate features, such as ASR where the feature vector has usually a double-digit dimension, the method based on mutual information has a practical drawback: as the dimensionality of the pattern (or feature vector) increases (e.g. up to 20 in the present study) the size of the necessary data and the computation time increase exponentially [8]. This makes the computation of the JMI feasible only between very few features (2 in the above studies). A solution to overcome this difficulty is to independently compute the MI between each feature and the phoneme identity, but this neglects the relationships among the elements of the feature vector which are important for classification. Also, the average MI over  $D$  features and the JMI of the  $D$  features are not the same. For this reason, this study employs a direct method based on classification scores instead of mutual information to estimate the underlying distribution of information correlated with vowel identity. The classification method can account for the relationship among features, but it is sensitive to the actual probability distributions of the features when these are approximated by parametric models

such as multivariate Gaussians. However, it should be noted that the classification scores and the JMI are not identical and the former depend on the classification method used.

The analysis and classification are done for eight types of patterns. The details of the first seven types of patterns and their temporal positions along biphones can be found in [6]. In addition to these patterns another pattern was included in this study to expand the limits of the analysis of the distribution of information. Briefly, the eight patterns are described as follows:

1. Spectral feature vector at the center of the vowel in CV or VC biphones.
2. Spectral feature vector at 20 ms after vowel onset in CV biphones or 20 ms before vowel offset in VC biphones.
3. Spectral feature vector at the CV or VC transition position.
4. A vector containing the overall slope of each spectral feature computed on a 40 ms interval, centered at the CV or VC transition position.
5. A vector containing the slopes of each spectral feature computed on 20 ms intervals on the left- and on the right-side of the given CV or VC transition position. This vector can discriminate among the monotonic and non-monotonic spectral transitions between phonemes as presented in [6] and [12].
6. Spectral feature vector at the center of the preceding consonant in CV biphones or the following consonant in VC biphones.
7. A vector containing the vowel duration and the duration of the consonant in CV or VC biphones. This vector accounts for both the intrinsic duration of vowels and the vowel durational effect due to coarticulation with consonants.
8. Spectral feature vector at the beginning of the consonant in CV biphones or at the end of the consonant in VC biphones.

Among the eight acoustic patterns evaluated in this study five (1, 2, 3, 6, and 8) represent static spectral features at various temporal positions along the CV or VC biphones, two (4 and 5) are dynamical spectral features, and one (7) represents temporal (durational) acoustic features. This study employs the maximum likelihood (ML) classification method which is applied to each of the eight types of patterns. Other classification methods, such as those based on neural networks, can also be applied but the general findings are expected to be similar. The classification of the nine vowels is done separately in each of the three-left (CV) and three-right (VC) consonant contexts and then the results are averaged across the left-consonant context (L) or right-consonant context (R). Thus, for example, classification scores denoted 3L represent the result of classification based on the spectral vector at the CV transition averaged across the three left-consonant contexts (/b/\_V, /sh/\_V, and /m/\_V). The spectral features used in this study are the Mel-

Frequency Cepstrum Coefficients (MFCC), as described in [13], because they are among the most common acoustic features used in ASR. For each spectral frame the first 10 MFCC features (excluding the MFCC<sub>0</sub> which represents the total energy) are computed on Hamming windows of 32 ms, with a frame step of 10 ms. The ML classification selects the vowel  $v^*$  based on the equation

$$v^* = \arg \max_v P(v|x) = \arg \max_v P(x|v)P(v), \quad (1)$$

where  $P(v|x)$  is the conditional probability of vowel  $v$  given the feature vector  $x$ ,  $P(x|v)$  is the feature probability given the vowel, and  $P(v)$  is the *a priori* probability of occurrence for vowel  $v$ . Modeling the probability distribution of the feature vector corresponding to vowel  $v$  by a multivariate Gaussian distribution function described as

$$f(x) = \frac{1}{(2\pi)^{D/2} |C_v|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_v)^T C_v^{-1} (x - \mu_v)\right\}, \quad (2)$$

where  $D$  is the dimension of the feature vector,  $\mu_v$  is the mean, and  $C_v$  is the covariance matrix of the feature vector  $x$  for vowel  $v$ , the ML classification can be reduced to

$$v^* = \arg \min_v \{[x - \mu_v]^T C_v^{-1} [x - \mu_v] + \ln |C_v|\}. \quad (3)$$

This classification formula is not optimal in a Bayesian sense because the underlying probability distributions of the features are not perfectly multivariate Gaussians and because it does not include the *a priori* probabilities of vowels. An alternative classification method which is optimal in the Bayesian sense can employ artificial neural networks.

For each of the eight types of patterns in each consonant context multivariate Gaussian models are built from the training corpus for each of the nine vowels. These models consist of the means and covariance matrices of the distributions of these features. Since the purpose of this study is to analyze the distribution of the acoustic-phonetic information in the given data along various acoustic patterns the vowel classification is not achieved on a separate test corpus but on the same training corpus as described in Section 2. The classification experiments are performed using both full- and diagonal-covariance matrices. In the case of full-covariance matrices, if the number of training samples for a particular model is less than the dimension of the feature vector then the diagonal covariance matrix is computed instead to avoid singularity in matrix inversion. However, there are only a few such cases in these data. The dimension of the feature vectors is: 10 for patterns 1, 2, 3, 4, 6, and 8; 20 for pattern 5, and 2 for pattern 7. Confusion matrices are computed for each type of pattern and consonant context. The correct classification score is computed as the ratio of the sum of the diagonal elements over the total number of elements in each confusion matrix

for each context. Then the classification results are averaged separately across the three types of left-consonant context or three types of right-consonant contexts.

#### 4. DISTRIBUTION OF INFORMATION: RESULTS

Figure 1 presents the vowel classification results for the eight types of patterns, averaged across the three left-consonant contexts (L) and Figure 2 presents the results averaged across the three right-consonant contexts (R). Temporal order of patterns is reversed in the two contexts. These figures present the classification results using full-covariances (Full) and diagonal-covariances (Diagonal). The scores displayed in red are based on full-covariances and in blue on diagonal-covariances.

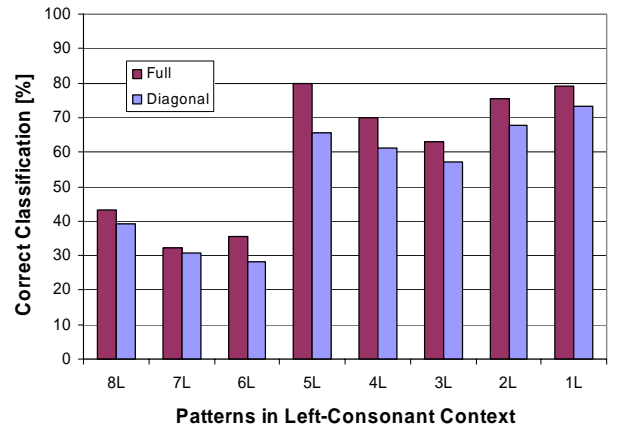


Fig. 1. Vowel classification results averaged across the three left-consonant contexts.

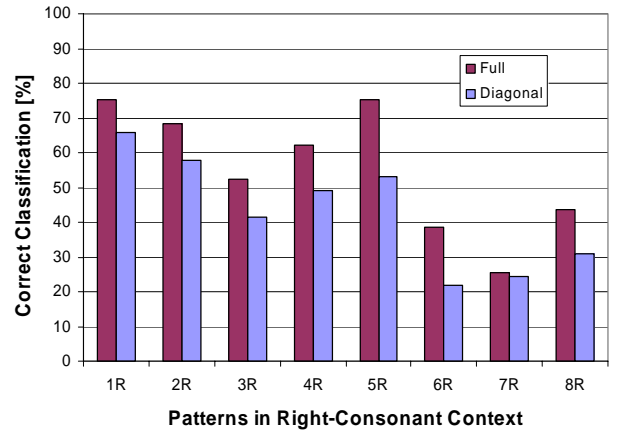


Fig. 2. Vowel classification results averaged across the three right-consonant contexts.

In analyzing these results it should be noted that the correct classification accuracy by selecting one of the nine vowels by chance is 11.1%. As expected, the scores based on full-covariances are higher than those based on diagonal-covariances. This is because the JMI between features is

higher than the average MI across features. All of the eight acoustic patterns evaluated in this study contain useful information about the vowel identity since their vowel classification accuracies are much higher (at least double) than the chance value, even in the diagonal-covariance case. As expected, the vowel classification score for the static MFCC patterns (1, 2, 3, 6, and 8) in both left (L) and right (R) contexts decreases when the position of the pattern relative to center of the vowel increases, but there is an exception: the 8<sup>th</sup> patterns that score better than the 6<sup>th</sup> patterns. This is somehow surprising. This decrease in classification scores (information) may be explained by a weaker vowel coarticulation effect in the middle of the adjacent consonant. The static MFCC patterns can be viewed as corresponding to 5 different states in an Hidden Markov Model (HMM). Another interesting characteristic is that for both full- and diagonal-covariance cases significant scores are achieved for the patterns outside of the provided vowel boundaries in TIMIT. This confirms the previous results that significant information for phoneme identity is distributed over an interval longer than about 100 ms on either side of the center of the phoneme (e.g. [9], [11]), but, unlike the previous studies, it quantifies this information with respect to the boundaries of the adjacent consonants. Another interesting characteristic is that the classification scores based on the dynamic MFCC patterns are higher for the patterns 5 than for the patterns 4. This means that the double-slope patterns 5 contain more information than the single-slope patterns 4, and supports the importance of the non-monotonic spectral transitional features introduced in [12] in phoneme classification. As shown by the two figures the patterns 5 achieved the highest scores (corresponding to the full-covariance case), higher or similar with those of patterns 1 from vowel centers. Also, the two-dimensional duration patterns 7 achieved scores double or triple the chance values. Yet, another observation is that the scores in the left- and right-consonant contexts are somehow asymmetrical. This asymmetry could be an intrinsic property of speech but it could also be due to differences in the data used in the two cases in this study. However, similar asymmetries in the distribution of information were found in [9] and [10] based on JMI, and in [11] based on MI. The slight asymmetry suggests that there is more information about the vowel identity in the region before its center than in the region after its center. The detailed analysis of this asymmetry is outside the scope of this paper.

Figure 3 presents the distributions of the vowel classification scores based on the static spectral features at the analyzed temporal positions in both left- and right-consonant contexts. Each curve contains in fact two separate curves, corresponding to the left-consonant context (8L to 1L) and to the right-consonant context (1R to 8R). It should not be viewed as a continuous distribution corresponding to CVC syllables. Thus, the line between the patterns 1L and 1R should be ignored since these patterns represent the

same temporal position at the center of the vowels but in two different types of context.

An analysis is performed to reveal if the vowel information in the new patterns (5 to 8) is significant and uncorrelated with the vowel information in patterns 1 to 4.

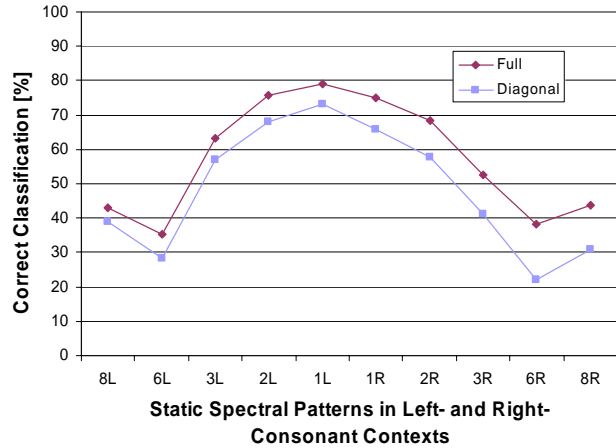


Fig. 3. Vowel classification scores using static MFCC patterns at various positions along CV and VC biphones.

Figure 4 presents the classification scores obtained by multiplying the classification probabilities of various static MFCC patterns. Due to limited space only the scores in the left-consonant context are analyzed in the rest of the paper. The increase in these scores by using more patterns is almost linear within and outside phoneme boundaries.

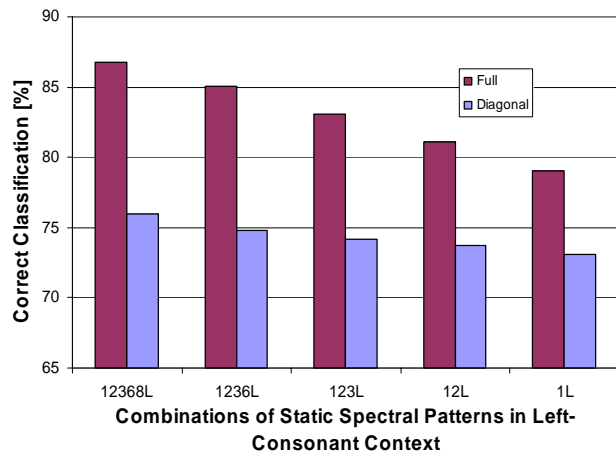
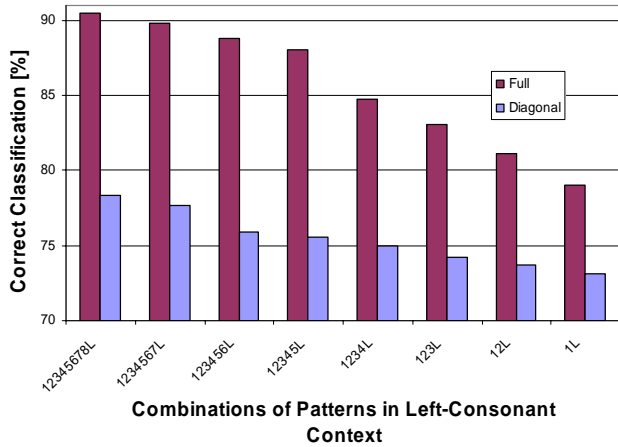


Fig. 4. Vowel classification scores based on linear combinations of static MFCC patterns.

Similarly, Figure 5 presents the classification scores obtained by linear combinations of all the 8 patterns in the left-consonant context. For the full-covariance case note the score increase from 84.5% to 90.5% (approx. 39% relative error reduction) by adding patterns 5 to 8. This means that the new patterns contain significant vowel information, uncorrelated with that in patterns 1 to 4. These patterns are

not directly used in the current ASR technologies, such as those based on Hidden Markov Models (HMM), for computing the probability of the phone models (either context-independent or context-dependent).

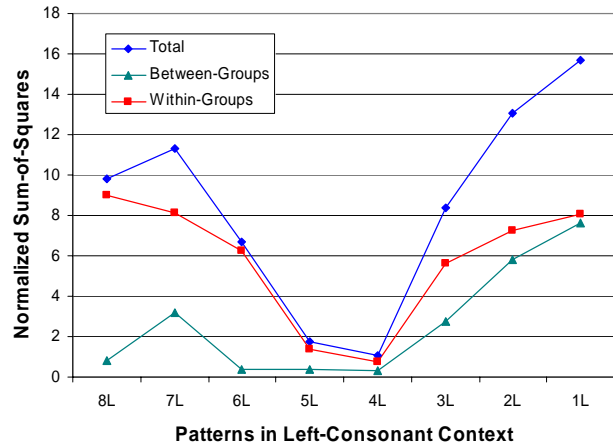


**Fig. 5.** Vowel classification scores based on linear combinations of all 8 patterns.

## 5. ANALYSIS OF THE INTRINSIC VARIABILITY: METHOD AND RESULTS

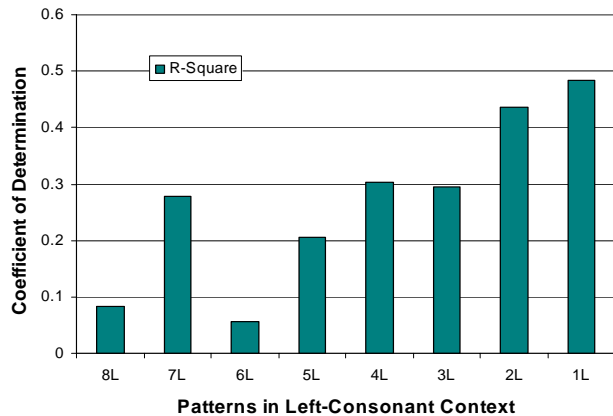
An analysis of the sources of the intrinsic variability of speech is performed to see how the results of this analysis correlate with and predict the shape of the underlying distribution of information correlated with vowel identity. This analysis is based on the Multivariate Analysis of Variance (MANOVA) statistical method. This analysis focuses on the interplay between the within-groups and between-groups sources of variability and it is significant for understanding such relationships and for developing new acoustic features for ASR. Only the data corresponding to the left-consonant context are used in this analysis due to limited space. MANOVA is performed for each of the three left-consonant contexts and then the results are averaged. The normalized Sum-of-Squares (SS) of this analysis are presented in Figure 6 for the total, between-groups, and within-groups variability for each pattern. The groups in this analysis are the 9 vowel types. Each SS value is computed here as the trace of the Sum-of-Squares-and-Cross-Product (SSCP) matrix from MANOVA. The normalization is done in this study by dividing the SS values from MANOVA by the number of features in each pattern and optionally by an arbitrary number which here equals the degree of freedom of the total variability. It is expected that greater between-groups variability and smaller within-groups variability correlate with higher classification scores. The curves show the shape of the distributions of these variabilities per-feature across the 8 patterns. The correlation between these distributions and the classification scores is not easily quantified by looking at these curves but it is better seen in an analysis of the coefficient of

determination, denoted  $R^2$  or R-Square, which is the ratio between the between-groups SS and the total SS from MANOVA.



**Fig. 6.** Normalized total, between-groups, and within-groups variability for patterns in the left-consonant context.

Figure 7 shows the distribution of the coefficient of determination across the 8 patterns in left-consonant context.



**Fig. 7.** Distribution of the coefficient of determination (R-Square) across the 8 patterns in the left-consonant context.

The R-Square is a measure which shows what proportion of the total variability is due to the existence of the groups (here the 9 vowels). Note some similarities between the R-Square distribution and the classification distribution for the diagonal-covariance case presented in Figure 1. However, there are important dissimilarities between the two distributions because the former represent a measure which does not reflect the dimension  $D$  of the feature vectors (it can be considered as an average per-feature measure) whereas the latter represent a per-vector (or per-pattern) measure which usually increases when the feature vector dimension increases. A better comparison could be made between the two distributions if the R-Square

measure is transformed to reflect the dimension of the feature vector. However, no claim is made that such a *ad hoc*, transformed R-Square has the same statistical characteristics as the R-Square measure. It is proposed here just to predict the per-pattern classification scores by analyzing the intrinsic variability. Figure 8 presents a comparison between the classification probabilities (not in percentage) for the diagonal-covariance case and the transformed R-Square which is called here normalized R-Square. The formula used to compute the normalized R-Square is  $\text{NormR}^2 = R^2 * (D/10)$ , where  $D$  is the dimension of the feature vector in each of the 8 patterns. Such normalization leaves the R-Square values for the patterns 1, 2, 3, 4, 6, and 8 unchanged and only adjusts the values of the pattern 5 and 7 which have dimensions 20 and 2, respectively. Linearly fitting the normalized R-Square measure ( $r$ ) displayed in Figure 8 (green bars) to the classification scores based on diagonal-covariance matrices (blue bars) results in the following linear predictor:  $y = 0.98r + 0.27$  (yellow bars) which predicts the classification scores with less than 5% error.

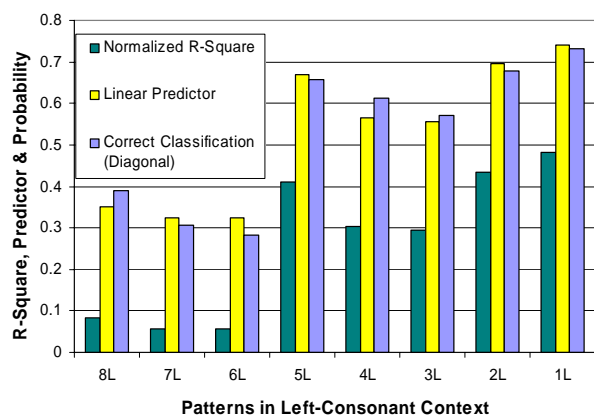


Fig. 8. Comparison between the normalized R-Square and the classification probabilities for the 8 patterns.

## 6. CONCLUSIONS

This study analyzed two different methods that may be useful in designing new ASR features. The first part of this study performs an analysis of the relevance of various acoustic patterns for classification of coarticulated vowels using a direct method (ML classification) instead of employing an indirect method based on information theory and indirect measures such as MI or JMI. The goal of this analysis was neither to obtain the best possible vowel classification score on any individual pattern nor to select the best features in a large concatenated vector like in [8], [9], and [11], but to reveal the relative amount of vowel information contained in patterns not currently used in ASR and in feature vectors which lie outside the currently accepted phoneme boundaries. Significant and uncorrelated

information has been found in the new patterns (shown by approx. 39% relative reduction in total classification error obtained by using patterns 5 to 8 in addition to patterns 1 to 4). The second part of this study proposes a new, accurate method to reveal the distribution of the vowel information across various patterns by employing an analysis of the intrinsic variability. Such simple method could be effective in evaluating new features for ASR. Current and future research focuses on integrating the proposed features into a full HMM system.

## 6. ACKNOWLEDGMENTS

The author thanks Lawrence Rabiner and Aaron Rosenberg for their comments on a previous version of this paper.

## 7. REFERENCES

- [1] G.E. Peterson, and H.L. Barney, "Control Method Used in a Study of Vowels", *J. Acoust. Soc. Amer.*, 24(2): 175-184, 1952.
- [2] B.E.F. Lindblom, and M. Studdert-Kennedy, "On the Role of Formant Transitions in Vowel Recognition", *J. Acoust. Soc. Amer.*, 42(4): 830-843, 1967.
- [3] W. Strange, R.R. Verbrugge, D.P. Shankweiler, and T.R. Edman, "Consonant Environment Specifies Vowel Identity", *J. Acoust. Soc. Amer.*, 60(1): 213-224, 1976.
- [4] W. Strange, "Dynamic Specification of Coarticulated Vowels Spoken in Sentence Context", *J. Acoust. Soc. Amer.*, 85(5): 2135-2153, 1989.
- [5] S. Furui, "On the Role of Spectral Transition for Speech Perception", *J. Acoust. Soc. Amer.*, 80(4): 1016-1025, 1986.
- [6] S. Dusan, "On the Nature of Acoustic Information in Identification of Coarticulated Vowels," *Proc. of INTERSPEECH/EUROSPPEECH 2005*, Lisbon, Portugal, 2005.
- [7] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D. S. Pallett, and N.L. Dahlgren, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.
- [8] A. Morris, J.-L. Schwartz, and P. Escudier, "An Information Theoretical Investigation into the Distribution of Phonetic Information Across the Auditory Spectrogram," *Computer Speech and Language*, 2: 121-136, 1993.
- [9] H.H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, "Relevance of Time-Frequency Features for Phonetic and Speaker-Channel Classification," *Speech Communication*, 31: 35-50, 2000.
- [10] K. Kirchhoff and J.A. Bilmes, "Statistical Acoustic Indications of Coarticulation," *Proc. of ICPHS 1999*, pp. 1729-1732, San Francisco, 1999.
- [11] P. Scanlon, D.P.W. Ellis, and R. Reilly, "Using Mutual Information to Design Class-Specific Phone Recognizers," *Proc. of INTERSPEECH/EUROSPPEECH 2003*, Geneva, 2003.
- [12] S. Dusan, "Non-Monotonic Spectral Transitions between Successive Phonemes," *J. Acoust. Soc. Amer.*, 116(4), Pt. 2, p 2479(A), 2004.
- [13] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition," *IEEE Trans. on ASSP*, 28 (4), 357-366, 1980.