STATISTICAL ACOUSTIC INDICATIONS OF COARTICULATION

Katrin Kirchhoff* and Jeff A. Bilmes†
*Applied Computer Science Group, University of Bielefeld, Germany
† ICSI/U.C. Berkeley, Dept. of EECS, CS Division, Berkeley, USA

ABSTRACT

Coarticulation in speech is one of the most difficult problems for automatic speech recognition (ASR) systems. The degree of coarticulation is assumed to vary with contextual conditions, such as differences in speaking rate, stress, etc. In the past, coarticulation has been studied using only limited data sets and using acoustic-phonetic methods such as formant analysis. We propose a method that statistically analyzes the degree of coarticulatory influence on features typically used for automatic speech recognition systems (LPCs, MFCCs, RASTA, and compressed subband spectral envelopes). This method computes the Conditional Mutual Information (CMI) between time/feature-position pairs under a variety of coarticulatory conditions. We applied this method on a two-hour subset of the Switchboard database and analyzed CMI for various speaking rate, stress, and vowel category conditions. Results show that CMI is indeed larger for those phonetic conditions believed to possess more coarticulation.

1. INTRODUCTION

Although automatic speech recognition has improved dramatically in recent years, widespread use of speech recognition devices is still far from a reality. One of the major shortcomings of existing speech recognizer systems (as identified by [5] for instance) is their limited capability to handle the coarticulation that exists in everyday conversational speech. Coarticulation is usually defined as a change in the acoustic-phonetic content of a speech segment due to anticipation or preservation of adjacent segments. Current techniques for statistical modeling used by speech recognition systems (e.g., context-dependent phone models) are believed to be insufficient for capturing all coarticulatory effects.

Phonetic research has identified a range of linguistic and extralinguistic conditions which affect the degree of coarticulation. Although most of these studies [4, 3, 7] are based on articulatory data, it is widely assumed that corresponding acoustic effects can be observed in the speech signal and that, if these conditions are modeled in an ASR system, performance of the system can be improved. Acoustic studies of coarticulation have so far concentrated on formant analysis (e.g. [8]); we are unaware of any quantitative statistical analysis on a large speech corpus which investigates the effects of coarticulation on the acoustic features typically used by speech recognizers (mel-frequency cepstral coefficients (MFCC), linear predictive coefficients (LPC), etc.).

In this paper we present a methodology which allows us to analyze the degree of influence of the surrounding acoustic context directly on speech features for a variety of coarticulatory conditions. Varying the coarticulatory condition, we observe the influence on acoustic features using the conditional mutual information (CMI). We investigate several coarticulatory conditions including: 1) speaking rate; 2) the degree of syllable stress (primary, secondary, or unstressed); and 3) vowel quality (central/lax vs. peripheral). Our assumptions are that faster speaking rate and lower degree of stress correlate with stronger coarticulation. Furthermore, we expect central and lax vowels to be more coarticulated than peripheral vowels and diphthongs.

2. METHOD

2.1. Conditional Mutual Information The entropy [1] of a random variable X is defined as

$$H(X) = -\sum_x p(x) \log p(x)$$

and can be interpreted as the average amount of uncertainty, or information, associated with the random variable X. The mutual information $I(X;Y)$ between two random variables X and Y, is defined as

$$I(X;Y) = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the variable Y and $H(Y|X)$ is the conditional entropy of the variable Y given the variable X. The mutual information $I(X;Y)$ measures the reduction in uncertainty that Y provides about X and vice versa, since mutual information is symmetric in its arguments. It thus gives an indication of the extent to which the variables X and Y are dependent, or in other words, how predictable Y is given knowledge of X.

The *Conditional* Mutual Information (CMI) $I(X;Y|Q)$ is the average entropy reduction of one variable X given knowledge of another variable Y conditioned on the knowledge provided by a third variable Q. That is,

$$\begin{aligned} I(X;Y|Q) &= H(Y|Q) - H(Y|X,Q) \\ &= \sum_q p(q) \sum_{x,y} p(x,y|q) \log \frac{p(x,y|q)}{p(x|q)p(y|q)} \\ &= \sum_q p(q) I(X;Y|Q=q) \end{aligned}$$

This last equation shows that CMI is the the average entropy reduction for different conditions, and that $I(X;Y|Q=q)$ is the entropy reduction for a particular condition q.

Feature type	num coeffs	energy	deltas	frame rate	win size
MFCC	12	yes	yes	12.5	25
RASTA	8	yes	yes	12.5	25
LPC	14	no	yes	12.5	25.6
CSSE	22	no	no	12.5	variable

Table 1: Details of different preprocessing methods used in CMI experiments. Frame rate and window size are given in milliseconds.

We compute CMI for a variety of different feature sets including mel-frequency cepstral coefficients (MFCCs), linear prediction coefficients (LPC), relative spectral coefficients (RASTA) [2], and cube-root-compressed sub-band spectral envelopes (CSSE). These features were computed for a randomly selected two-hour subset of the Switchboard corpus. Table 1 shows the details of each preprocessing method, viz. the feature type, number of basic coefficients, whether energy and/or delta coefficients were included, the frame rate, and the size of the analysis window.

For our present purpose, the X and Y variables correspond to a particular pair of feature elements from the three-dimensional space defined by the sequence of feature vectors. Let X_t be a sequence of feature vectors, and X_t^i be the i^{th} element of vector X_t. We consider pairs from the three dimensional set of the form $\{(X_t^i, X_{t-\ell}^j) : \ell, i, j\}$ for various "lags" ℓ, and feature element positions i and j. We first compute the CMI for each pair, and then summarize this information by computing the average over all i and j. This results in CMI plots as a function only of the (one dimensional) time-lag ℓ.

The condition $Q = q$ corresponds to the coarticulatory condition under investigation. This is done by labeling the speech frames with a particular coarticulatory condition, and for each case, computing the mutual information using only those frames matching the condition. The specific label sets used will be described separately for each experiment. All labels, however, were derived from the same phone transcriptions, which were obtained from an automatic alignment of the speech data using triphone models.

To provide a meaningful time unit where one coarticulatory condition can be compared with another, each individual CMI plot is time normalized to the average duration of that condition. Therefore, the abscissas are presented in sub-word units, where the sub-word unit *roughly* corresponds to a syllable duration (about 200ms) for the speaking rate and stress plots, and phone duration for the vowel category plots.

3. EXPERIMENTS

3.1. Speaking Rate It is usually assumed that fast speech exhibits more coarticulatory effects than slow speech. A faster speaking rate causes a larger degree of articulatory gesture overlap - as a consequence, acoustic feature vectors corresponding to fast speech should exhibit more inter-dependence. This should correspondingly cause an increase in CMI between individual feature vector components across time, compared to the more slowly articulated portions of the speech signals.

We use the quantitative procedure *mrate* to determine the speaking rate [6]. Mrate combines the results of **m**ultiple estimators of speaking **rate**, each of which is based exclusively on the acoustic signal. Therefore, no use is made of any lexical segment hypotheses. More specifically, the various estimators compute: 1) the spectral moment of a full-band energy envelope, 2) a peak count of the spectral energy envelope, and 3) a peak count of pointwise cross correlation of subband energy envelopes. The average of these measures, mrate, correlates fairly well[1] with the average number of syllables per segment and can thus be considered a reasonable indication of speaking rate. While a better speaking rate indicator is desirable, this simple measure seemed sufficient for our purposes.

The mrate values assigned to each frame in the two-hour dataset were quantized into three categories: slow, medium, and fast. The boundaries for these classes were placed at the 33rd and 66th percentiles, respectively, of the distribution of mrate values. As an additional category, 'silence' was also used, on which CMI was not computed. The leftmost graphs of Figures 1 through 4 plot the resulting CMI curves. On the vertical axis, these plots show CMI[2] and on the horizontal axis they show the temporal distance (in terms of roughly the average syllable length) from the current frame. The CMI values for the zero-lag point ("0") are omitted from the plots as these points express only the self-information [1] (or entropy) of a feature vector with itself and are irrelevant for the present study. As can be seen, for all types of speech features, a faster speaking rate induces a larger CMI, which, according to our assumptions, indicates greater coarticulation

3.2. Stress Another frequently encountered assumption is that stressed portions of the speech signal are not as heavily coarticulated as unstressed portions. In order to verify this assumption, we assigned stress labels to the signals in our data set. Since stress is a feature of the syllable nucleus, the data was segmented into syllables. Each syllable was then assigned one of three stress labels, primary stress, secondary stress, or unstressed, based on the word context and the word-based stress marks in the Pronlex dictionary. It should be noted that this labeling does not reflect stress as it is realized in the actual speech signal but a hypothesis of how it might be realized based on canonical lexical definitions. This approximation was used because signal-based stress detection algorithms are notoriously unreliable. The use of syllable-sized segments entails a further problem: long syllables may already include much of the phonetic context that influences the stressed syllable nucleus. For such cases, the CMI computed over a fixed range surrounding the entire syllable will not reflect the true coarticulatory effect on the nucleus. For this reason, long syllables were split into two parts, and CMI was computed separately for each of these cases. The splitting threshold was set at $E(S) - 1.5 * stddev(S)$ where $E(S)$ (resp. stddev(S)) is the mean (resp. standard deviation) of the syllable duration

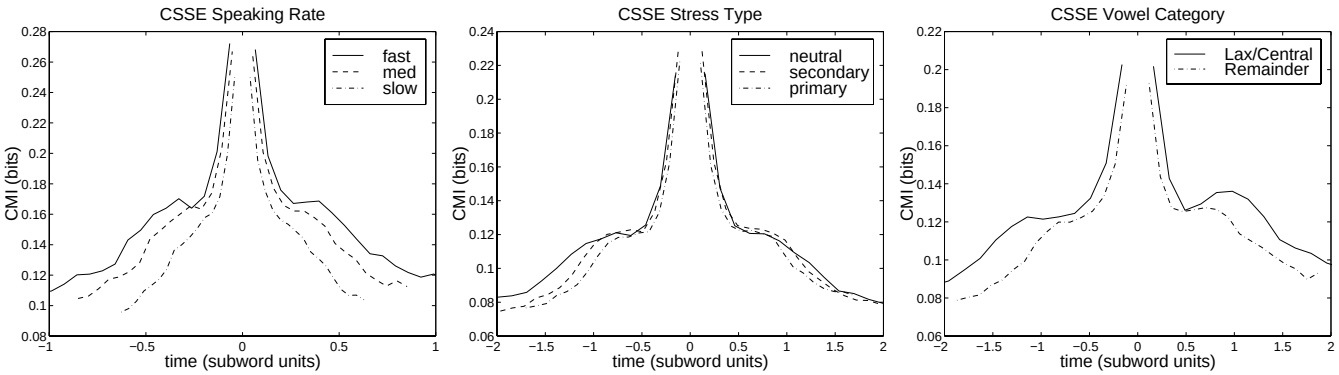

Figure 1: CMI for CSSE features. Speaking rate, stress, and vowel category.

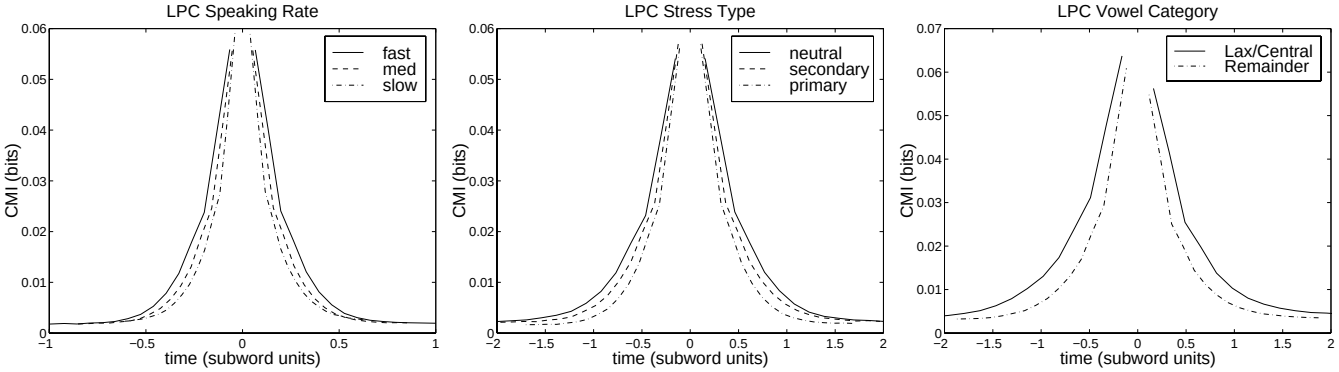

Figure 2: CMI for LPC features. Speaking rate, stress, and vowel category.

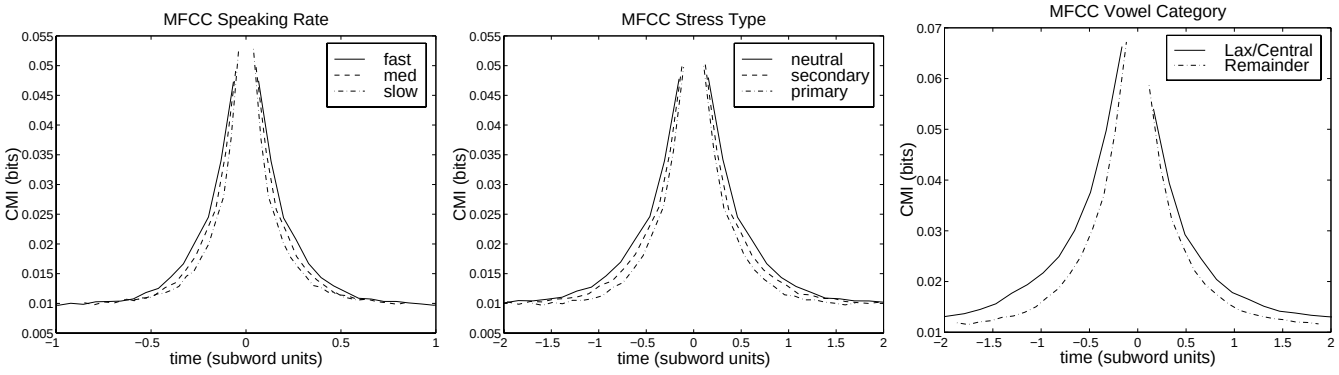

Figure 3: CMI for MFCC features. Speaking rate, stress, and vowel category.

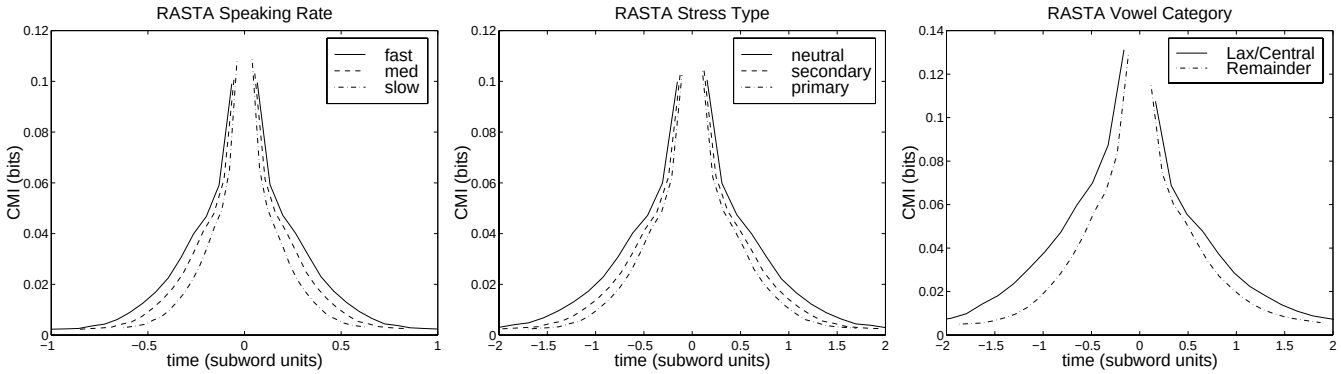

Figure 4: CMI for RASTA-PLP features. Speaking rate, stress, and vowel category.

as computed on the data set. Syllables whose duration fell below this threshold were not split; all other syllables were split in half. This led to 10 labels overall, three possible stress labels for short syllables (primary,secondary,unstressed) six possible labels for long syllables (primary-beginning, primary-end, secondary-beginning, secondary-end, unstressed-beginning and unstressed-end), and 'silence'.

The central graphs of Figures 1 through 4 show the resulting CMI curves. Again, we see our assumptions confirmed: CMI curves for neutral (unstressed) syllables exceed those for secondary stressed syllables, which in turn are higher than those for primary stressed syllables. This holds for all types of features.

3.3. Vowel Quality Finally, we investigated the assumption that central vowels such as schwa are more coarticulated than peripheral vowels and diphthongs. To this end, we divided the set of vowels in the original phoneme set into two classes (central and peripheral), as shown in the following table:

Central/Lax Vowels	Peripheral Vowels
ah,ax,ih,eh,uh	aa,ae,ao,aw,ay,eh,er,ey,iy,ow,uw

CMI values were computed separately for each category; the results are shown in the rightmost graphs of Figures 1 through 4. The plots, once again, confirm the existence of greater coarticulation for central and lax vowels.

4. DISCUSSION

For the various conditions typically assumed to increase coarticulation (high speaking rate, unstressed syllables, and central/lax vowels), we find a corresponding increase in the conditional mutual information. This analysis has potentially important implications for the front-end and acoustic model design of speech recognition systems as it indicates the degree to which acoustic realizations of feature vectors can be affected by contextual factors under different coarticulatory conditions. Therefore, a system that respects this acoustic contextual dependence might perform better on everyday conversational speech.

Acknowledgments

We thank Eric Fosler-Lussier for providing the mrate values, and Steve Greenberg for useful suggestions concerning this work. Katrin Kirchhoff was supported by the graduate program "Task-oriented communication" at the University of Bielefeld, Germany. Jeff Bilmes was supported by a DoD IDEA grant.

Notes

[1] The correlation coefficient is .6

[2] An empirical analysis was performed on the CSSE features to obtain a significance level. By computing CMI on Gaussian noise audio signals represented as CSSE features, it was found that a difference of greater than about 0.01 bits could be considered significant. For other features types, the significant difference would likely be smaller than 0.01 bits.

5. REFERENCES

1. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.
2. H. Hermansky and N. Morgan. Rasta-processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.
3. I. Hertrich and H. Ackermann. Coarticulation in slow speech: durational and spectral analysis. *Language and Speech*, 38:159, 1995.
4. P. Hoole, N. Nguzen-Trong, and W. Hardcastle. A comparative investigation of coarticulation in fricatives: electropalatographic, electromagnetic and acoustic data. *Language and Speech*, 36:263–288, 1993.
5. R. Lippmann. Speech recognition by humans and machines. *Speech Communication*, pages 1–15, 1997.
6. N. Morgan and E. Fosler-Lussier. Combining multiple estimates of speaking rate. *ICASSP 98*, pages 729–732, 1998.
7. K. van de Jong, M. Beckman, and J. Edwards. The interplay between prosodic structure and coarticulation. *Language and speech*, 36:197–212, 1993.
8. E.C. Zsiga. Acoustic evidence for gestural overlap in consonant sequences. *Journal of Phonetics*, 22:121–140, 1994.