INFORMATION THEORETIC ANALYSIS OF DIRECT ARTICULATORY MEASUREMENTS FOR PHONETIC DISCRIMINATION

Jorge Silva, Vivek Rangarajan, Viktor Rozgic and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, *http://sail.usc.edu*

University of Southern California, Viterbi School of Engineering

jorgesil,vrangara,rozgic@usc.edu, shri@sipi.usc.edu

ABSTRACT

This paper focuses on the analysis of speech production signals (physical measurements from electromagnetic articulograph) from the perspective of phone discrimination. We explore two different signal representation schemes for the articulatory signals, one based on time-domain analysis and the other based on frequency domain. We quantify the amount of discrimination information offered by the speech production signals in identifying the phone labels through mutual information. Mutual information analyses establish that substantial discrimination information is present in the articulatory stream. Furthermore, phonological classification results with articulatory signals indicate higher accuracy compared to the acoustic signal.

Index Terms— Analysis of articulatory measurements, automatic speech recognition, non-parametric mutual information estimation, phonological classification.

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) systems use parameterizations of the acoustic speech signal for recognition. However, this representation of speech does not consider the actual speech production process which results from continuous movement of articulators from one configuration to the next. These low frequency speech production signals characterize an intermediate representation of the phonetic labels typically associated with the acoustic signal. Hence, directly modeling the underlying process that generates the acoustic signal promises to provide more discriminatory information in recovering the accurate phonetic sequence.

Articulatory features (AFs) have been proposed for ASR mainly from two different perspectives. One, from the perspective of modeling coarticulation [1, 2, 3], which surmises that the joint modeling of acoustic-articulatory streams could better account for the co-articulatory effects and the other, from the robustness of AFs to environmental noise that its acoustic counterpart suffers from [4]. AFs can be obtained either from direct physical measurements of articulatory movements through techniques like magnetic resonance imaging (MRI), electromagnetic midsaggital articulography (EMA) [3] or as discrete knowledge-based representations which describe either articulation [1, 4] like *voicing, manner, place*, etc. or mechanics [2] like *tongue, jaw, lip positions*, etc. The drawback of the former acquisition method is that the direct observations are difficult to obtain and typically not available during recognition. Inverse mapping procedures using neural network [4] and HMM-based models [5] have been presented to predict the articulatory evidence from acoustic evidence but with limited success.

This material is based upon work supported by awards from NSF, ONR-MURI, and the U.S. Army.

AFs have been incorporated in several modeling schemes to jointly model the acoustic and articulatory evidence. However, there has been lack of a systematic study where the discriminatory capability of the articulatory features have been shown to aid or buttress the discrimination offered by the acoustic features in resolving ambiguity in phone classification. Furthermore, appropriate signal representations of the directly-obtained articulatory measurements from the motivation of phonetic discrimination have not been proposed. Another drawback of most articulatory feature based systems is that they depend on the speech segmentation for alignment. However, the articulatory and acoustic signals evolve at different rates and a synchronous frame-level analysis ignores the asynchrony between the two streams. For e.g., the articulatory evidence for a stop may occur much earlier in time than the corresponding acoustic evidence (phonetic symbol).

In this paper we present an information theoretic approach to investigate the dependencies between the discrete phone labels and the articulatory and acoustic speech streams. Our objective is two-fold, one is to explore appropriate signal representations for the speech production signals and the other is to evaluate if the speech production signals provide complementary information for speech recognition. We perform the analysis with the mutual information (MI) which is widely used as an objective metric in classification and estimation problems [6, 7]. We also present phonological classification experiments based on acoustic and articulatory representations that further corroborate the usefulness of modeling the speech production signals for recognition. Finally, we explore varying window lengths for articulatory signal representation, assessed in terms of MI to find the optimal analysis window length to capture the asynchrony between the acoustic and articulatory streams.

2. ARTICULATORY DATA ACQUISITION AND PROCESSING

We have recently collected a unique articulatory data set of *spontaneous speech* dialogs from a native American English speaker. In that regards, it is distinct from the MOCHA database [8] which provides EMA sensor measurements for read TIMIT sentences. We believe that the additional stream of information from articulatory data can be especially useful in modeling spontaneous speech and help compensate for the weakness of relying on the acoustic model alone.

We used the magnetometer point-tracking technology (EMA) to measure the articulatory movements. The EMA technique provides ideal temporal resolution for examining perturbations that occur at phrase junctures and under focal accent, offers well-established analytical approaches to speech kinematics, and is also accompanied by high-quality audio signal recordings. The articulatory data comprises two-dimensional continuous time signals, that capture the rel-

ative movement of the production of speech along different positions of the vocal tract. The articulatory sensors were placed on the subject's jaw (JAW), lower lip (LL), upper lip (UL), tongue tip (TT), tongue body (TBL) and tongue dorsum (TD). Acoustic and articulatory traces were simultaneously recorded at sampling rates of 200 Hz and 16 KHz, respectively. The data was subsequently corrected for head movements through rotation and translation to ensure that the two reference coils (upper incisor and bridge of the nose) were coincident across all frames for the speaker. A further rotation was performed to align the bite plane with the x-axis and a final translation to set origin at the upper incisor.

The speech data consists of *conversational-style* speech from a native American English (AE) male subject recorded in a dialog scenario. The recording sessions were moderated by an AE female interlocutor and the subject had no previous knowledge of the topics. 14 sessions on different topics were recorded, each lasting about 8 minutes. The audio of both sides of the conversation was recorded. The spontaneous speech dialogs were manually segmented into individual utterances and transcribed at the word level with appropriate disfluency tags. This process generated about 500 individual utterances. The speech signals were then aligned to the word level transcriptions using an ASR. The acoustic model used for this purpose was trained on 220 hours of spontaneous speech from the Fisher English Corpus, and adapted to the subject's speech through maximum likelihood linear regression (MLLR). The resulting phonetic segmentation of the speech signal obtained from the automatic forced alignment was used as a reference in the analysis of the production time-series signals (JAW, LL, UL, TT, TBL, TD). The next section presents two feature representation schemes for the raw speech production data used in subsequent analysis and classification.

3. BASIC ANALYSIS OF SIGNAL REPRESENTATION FOR CLASSIFICATION

We consider two signal representation schemes, one a time domain analysis, similar to that recently presented in [3] using similar raw production data, and the second a frequency domain based on a uniform filter-bank analysis. We implemented a frame-by-frame approach for feature extraction where, for each phone segment, we generated a sequence of overlapping windows with 10ms frame period, and 20ms window length. This offers reasonable resolution in time to capture some of the coarticulation events in the production signals, consistent with previous studies [3].

3.1. Time Domain Analysis

The raw speech production signals measured from the EMA setup are time varying x- and y- coordinates of the articulators produced every 5ms. For the time domain analysis we considered the articulatory position, velocity and acceleration data for each frame period of 10ms. To ensure time synchronous pairs of acoustic and articulatory data, we used every second position vector. This is similar to the approach described in [3] for deriving the feature vectors. Thus, a 3-dimensional vector time series is generated per coordinate x- and y- for each production signal stream.

3.2. Frequency Domain Analysis

In this setting, we used wavelet packets (WP) iterating a Daubechies' two channel filter bank block (db4) with balanced full tree-structure [9], as a flexible way of implementing uniform filter bank analysis. We conducted this analysis for each x- and y-spatial coordinate of the production signals. Coefficients obtained from the wavelet packet analysis provide the highest possible frequency resolution for a 20ms analysis window, which is conceptually equivalent to an STFT type of analysis. Then, the energy of every coefficient was used as a feature, where given the window size in this setting, it generates a 4 dimensional vector (4 uniform bands) per coordinate x- and y- for each production signal stream.

3.3. Mutual Information Estimation

For measuring the level of statistical dependency between the articulatory feature vector $X(u)$ and the discrete phone label $Y(u)$, we consider the mutual information between $X(u)$ and $Y(u)$ as our fidelity criterion. MI presents a strong relationship with the probability of error of Bayes classification approaches, mainly because of Fano's inequality [10], which characterizes a lower bound on the probability of error for any decision framework that tries to infer $Y(u)$ as a function of $X(u)$ [10].

The analysis schemes presented in the previous sub-section result in one observation vector for each frame along with the corresponding phone label information. The observation vector is the concatenation of the features obtained from the different articulatory signal streams. This information is represented by $T = \{(x_i, y_i) : i = 1, .., n\}$, x_i taking values in $\mathbb{R}^K$ and y_i in a discrete alphabet denoted by $\mathcal{A}_y$. Given that we do not explicitly have access to the joint distribution $P_{X,Y}$, but instead have a family of iid realizations T, we need to estimate the MI based on T.

We address the MI estimation by quantizing the feature observation space with a finite number of quantization bins, then estimating the observation-class distribution in the newly quantized finite alphabet space using standard maximum likelihood criterion — frequency counts — [11]; and finally applying the discrete version of the MI [10]. More precisely, let us denote $Q(\cdot) : \mathbb{R}^N \to \mathcal{A}_x$, where $|\mathcal{A}_x| < \infty$, is the quantized function, then the MI is given by:

$$I(Q(X), Y) = \sum_{q \in \mathcal{A}_x, y \in \mathcal{A}_y} P(Q(X) = q, Y = y) \cdot \log \frac{P(Q(X) = q, Y = y)}{P(Q(X) = q) P(Y = y)}, \quad (1)$$

where $Q(X)(u) = Q(X(u))$ is the quantized observation random variable. It is well known that $I(Q(X), Y) \leq I(X, Y)$, because quantization reduces the level of dependency between random variables. On the other hand, cleverly increasing the resolution of $Q(\cdot)$, implies that $I(Q(X), Y)$ converges to $I(X, Y)$ as the number of bins tends to infinity [12]. However, this result assumes that we know the joint class-observation distribution, which implies having an infinite amount of training data and a consistent learning approach. Consequently, for the finite training data scenario there is a tradeoff between how precisely we want to estimate $I(Q(X), Y)$, versus how close we want to be to the analytical upper bound $I(X, Y)$. We decided to have a resolution of $Q(\cdot)$ that guarantees good estimation of the joint observation-class distribution, and consequently a precise lower bound estimation for $I(X, Y)$. K-means vector quantization was used to characterize the quantization mapping [11, 10]. K-means is designed to minimize the mean square error between $Q(X)(u)$ and $X(u)$, for a given number of quantization bins, and consequently, is a good adaptive way of performing this mapping.

3.4. Experiments: Mutual Information Analysis

In this section, we calculate the MI between each of the oral articulators[1] and phone labels using the framework described above.

[1] Evaluating MI for all the signals together leads to data sparsity due to the large dimensionality of the resulting feature vector

The quantized MI, Eq.(1), is calculated for each articulatory signal (JAW, LL, UL, TT, TBL, TD), considering the two types of analyses presented previously. The entire dataset consisting of 500 utterances and corresponding to about 1 hour of speech-production signals, was used for the observation-class probability estimation. For each production signal stream, K-means vector quantization with 500 prototypes was used for both the 8-dimensional observation vector, in the filter-bank case, and the 6 dimensional observation vector, in the time-domain analysis. More than 200,000 realizations of the observation-class random vector $(Q(X)(u), Y(u))$ were used for having good empirical estimation of $\{P(Q(X) = q, Y = y) : 1 \leq q \leq 500, y \in \mathcal{A}_y\}$.

The same type of analysis was conducted for the speech signal, in order to have a reference value for the MI objective indicator. For this, we considered 13-MFCCs for the 20ms window length and K-means with same number of quantization bins. Delta and acceleration coefficients were not considered, because these implicitly introduce information akin to a longer window context. Hence, for a fair comparison we decided to restrict it to similar analysis (20ms window length, 10ms frame period) as that performed for the articulatory signals.

	filter bank	time-based
JAW	0.170281	0.342324
UL	0.237575	0.399443
LL	0.268569	0.455060
TD	0.362222	0.628452
TT	0.332358	0.669006
TBL	0.351672	0.721166
speech	1.272	n/a

Table 1. Mutual information of the filter bank and time-based analyses with respect to the phone label per production signal

Table 1 presents estimated MI values for the filter-bank and the time-domain analyses. In particular, in both the time domain and the filter-bank representations, TD, TT and TBL trajectories — production signals capturing the dynamics of the tongue — clearly demonstrate higher dependency with the phone-class information than JAW, UL and LL signals. This can be expected as the tongue movements are strongly correlated with vowels [1].

The time-domain analysis turns out to be more effective in capturing the relevant discrimination information provided by the production signals. One of the reasons for the filter-bank approach not performing as well for this analysis (restricted window length) is that, it only considers energy of the signals in the frequency bands and, consequently, it is a lossy representation of the original production signals. In contrast, for the time-domain analysis it is possible to recover the original signal from the time-based signal representation. It is important to note that these results are valid assuming a 20ms analysis window. This window size limits the level of frequency resolution obtained for the filter-bank representation, and also the level of long-term dependency obtained from the production signals in the time domain. This issue is addressed in more detail in Section 4.

The mutual information considering the whole 13-dimensional MFCC feature vector is 1.272. This value is greater than the MI of any single point-tracking production signal, which is reasonable considering that individual articulators provide only partial production information. However, it is notable that the MI of the production signals in Table 1 are still significant relative to the acoustic-phone MI: in particular, the time domain analysis has values of MI across each of the different production signals in the range of [31% - 57%] relative to the acoustic speech MI.

These results suggest that the oral production signals from EMA carry valuable phone discrimination information albeit not to the extent of the acoustic feature vector. In this direction, it is important to investigate if the level of MI present in each of the oral articulators translates to an improvement in overall phone classification accuracy. This is the focus of the next section.

3.5. Experiments: Phonological Factor Classification

In this section we evaluate classification performances using the direct production data with respect to the acoustic speech data, for categorizing different phonological (or articulatory feature) classes. For these experiments we use the time domain representation [3] based on our previous MI analysis.

Phonological classes are production oriented descriptions of phonetic units. They provide a way to cluster the phonetic labels based on gross-discretized description of the production formation process in the vocal tract. We consider manner (6 classes), place (10 classes), front-back (3 classes) and rounding (4 classes) as categories for the classification task (Table 2). From a practical point of view, these tasks are simpler than phone classification, as we can better address the sparseness of training data (we have just 1 hours of data). In considering these global categorization classes we have less estimation error effects and hence get more reliable information from the data, which was a critical issue that we had to address during all the analysis.

Phonological Classes	Phonological labels
Manner	vowel, lateral, nasal,fricative, approximant,silence
Place	dental, coronal, labial,retroflex, velar,glottal, high, mid, low, silence
Front-back	front, back, nil, silence
Rounding	+round, -round, nil, silence

Table 2. Phonological classes

The dataset was partitioned into training and test sets, 466 utterances for the training part and 37 utterances for the testing part (1 session). Class label for every category was obtained based on the phone-level transcription of the data and the canonical rule-based mapping from phone label to phonological factor. 36 dimensional time domain feature vectors were created for every frame by concatenating information for all the articulators [3]. Hidden Markov models (HMM) were trained using HTK 3.0 and standard EM-learning algorithms. Context independent HMMs with 3 observable states and 16 mixture components per state were used for both the production and acoustic features. We compare results with respect to the speech signal using the 13-MFCCs, in order to have the same window length (20ms) in both scenarios.

Performance for the different phonological factors are presented in Table 3. From these results it is evident that the production signals demonstrate better performance than their acoustic counterparts for almost all the phonological factor classifications (manner, place and font back) and very close to speech performance for the rounding task. Overall from the analysis of the type of errors incurred on the production side, it is interesting to note that direct speech production measurement had significantly fewer false alarm events than the speech side, in terms of insertion, across all the dimensions of classification explored. Consequently, based on these results, decisions based on the direct-measurements on average can be considered more reliable than decisions based on the acoustic information. However, the articulatory signals exhibit higher rate of deletions, primarily due to the inability of the short analysis window to capture the long-term dependencies. The results offer further evidence of the complementary nature of the production measurements for phone classification.

		ACC	C	D	S	I
Manner	production	45.34	899	612	249	101
	speech	35.11	1185	229	346	567
Place	production	46.42	974	509	275	158
	speech	28.24	1172	181	407	675
Front-back	production	57.27	1075	553	132	67
	speech	53.69	972	638	150	27
Rounding	production	52.08	933	749	73	19
	speech	57.84	1063	580	117	45

Table 3. Phonological factor classification performance using acoustic and direct production data. C: correct, I: insertions, D: deletion, S: substitutions and ACC: classification accuracy ($100 * \frac{(C-I)}{N}$)

It is important to mention that the results presented thus far are restricted to frame-by-frame type of analysis with fixed window length (20ms), which is far from being the optimal way of representing the production signals for phone classification. Consequently, this preliminary analysis is positively biased toward the acoustic side considering that the forced alignment was made based on it for the mutual information computation. To maximize the discrimination benefit offered by the direct production data, the frame-by-frame approach needs to be relaxed and one needs to explore the long-term dependencies of the articulators that would be more relevant for the phone classification task. We present important preliminary results for addressing this issue in the following section.

4. ANALYSIS OF TIME LOCALIZATION OF ARTICULATORY SIGNALS

The 20ms window length chosen in previous experiments is somewhat arbitrary for articulatory data analysis and relevant long-term dependencies between the articulators are not taken into account. By relaxing this constraint, the idea is to find a reasonable window of analysis that captures on average, significant information related to the acoustic-phonetic task without compromising much on the "complexity" of the problem, in terms of dimensionality. Again mutual information is used as a fidelity discrimination indicator for investigating the quality of the feature space.

In order to evaluate time localization properties, we consider raw temporal data at different window lengths (20ms, 40ms, 80ms and 160ms) in the 10ms frame rate setting. For this raw data, dimensionality reduction was computed based on principal component analysis (PCA) [11], which generates observation vector of same dimension across the different window sizes. Mutual information was estimated for all the window dependent feature vectors using the previously presented non-parametric approach, Section 3.3. Ensuring the same dimension and amount of training data is crucial for a fair comparison based on our estimation of the mutual information. The results are presented in Figure 1.

From Figure 1, one can infer that there is a significant gain in MI from 20ms to 40ms and also an important improvement from 40ms to 80ms; however beyond 80ms the fidelity gain is marginal [2]. The results show the same relative behavior across all production signals in the analysis. Note that the relative importance of the articulators in terms of mutual information is consistent with results presented in Section 3.4. Finally, the obtained optimal time range (80ms-100ms) brings up the problem of optimal signal representation, because we are now in a scenario where the tradeoff between feature complexity and representation quality needs to be addressed.

[2] 80-160ms is comparable to the time range used for the speech signal, considering the effect of delta and acceleration coefficients.

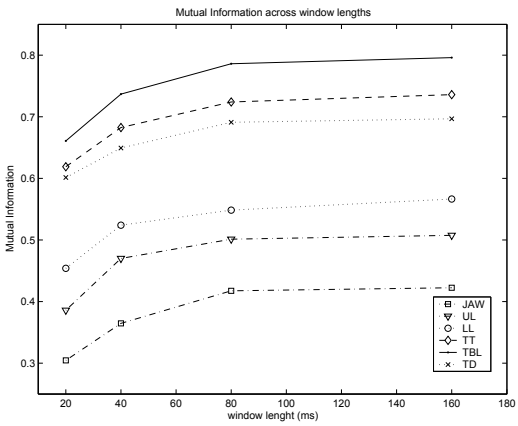

Fig. 1. Mutual Information across different window lengths for all the different production signals.

5. CONCLUSIONS

We presented an information theoretic analysis of articulatory signals collected via electromagnetic articulograph (EMA) for the problem of phonetic classification. The mutual information analysis show that EMA articulatory measurements provide significant phone discrimination information relative to the acoustic speech signal. We further corroborated the complementary nature of speech production signals by performing phonological classification. Finally, we explored the effect of varying the analysis window length to capture the long-term dependencies of the articulators relevant for phone discrimination. We are currently addressing the problem of dimensionality reduction and optimality of feature representation for longer window sizes (80-100ms). We are also working on schemes to jointly model the asynchronous feature streams using graphical models.

6. REFERENCES

[1] K. Erler and G. H. Freeman, "An HMM-based speech recognizer using overlapping articulatory features," *The J. of Ac. Soc. of Am.*, vol. 100, no. 4, pp. 2500–2513, October 1996.

[2] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator markov models for speech recognition," *Speech Comm.*, vol. 41, pp. 511–529, 2003.

[3] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Comm.*, vol. 48, pp. 161–175, 2006.

[4] K. Kirchoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Comm.*, vol. 37, pp. 303–319, 2002.

[5] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Tran. on Sp. and Audio Proc.*, vol. 12, no. 2, pp. 175–185, March 2004.

[6] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. on Image Proc.*, vol. 10, no. 11, pp. 1647–1658, Nov 2001.

[7] A. Ihler, J. Fisher, and A. Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Trans. on Signal Proc.*, vol. 52, no. 8, pp. 2234–2249, August 2004.

[8] A.Wrench, "The mocha-timit articulatory database," Queen Margaret University College, Tech. Rep., 1999.

[9] M. Vetterli and J. Kovacevic, *Wavelet and Subband Coding*. Englewood Cliffs, NY: Prentice-Hall, 1995.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, New York, 1991.

[11] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1983.

[12] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partition of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.