ANALYSIS-BY-SYNTHESIS FEATURES FOR SPEECH RECOGNITION

Ziad Al Bawab[†], *Bhiksha Raj*[‡], *and Richard M. Stern*[†]

[†]Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA. 15213
{ziada, rms}@cs.cmu.edu

[‡]Mitsubishi Electric Research Labs
Cambridge, MA. 02139
bhiksha@merl.com

ABSTRACT

We present a framework for speech recognition that accounts for hidden articulatory information. We model the articulatory space using a codebook of articulatory configurations geometrically derived from EMA measurements available in the MOCHA database. The articulatory parameter set we derive is in the form of Maeda parameters. In turn, these parameters are used in a physiologically-motivated articulatory speech synthesizer based on the model by Sondhi and Schroeter. We use the distortion between the speech synthesized from each of the articulatory configurations and the original speech as features for recognition. We setup a segmented phoneme recognition task on the MOCHA database using Gaussian mixture models (GMMs). Improvements are achieved when combining the probability scores generated using the distortion features with the scores using acoustic features.

***Index Terms*—** Articulatory Synthesis, Articulatory Recognition

1. INTRODUCTION

The speech signal is produced by the excitation of the vocal tract whose configuration is largely governed by the positions and dynamics of various articulators. Speech signals are therefore restricted to the set of signals that can actually be produced by a valid configuration of the articulators – configurations that are naturally likely to occur during speech and which cannot be eliminated by physical considerations. Furthermore, the actual *sequence* of sounds that can be spoken is also restricted by the dynamics of the articulators, which are governed by physical properties such as their mass and inertia.

Although the relationship between the vocal tract (and thereby between articulatory configurations) and the speech signal that is produced is well established, this relationship is not explicitly utilized in state-of-the-art speech recognition systems. Instead, the speech signal is usually treated in an entirely phenomenological manner: features that are derived for speech recognition are based on measurements of the spectral and temporal characteristics of the speech signal [1] without reference to the actual physical mechanism that generates it. Even spectral estimation techniques such as linear predictive coding, that purportedly model the vocal tract, do not directly access the physics of the generating process – the relationship between the parameters that are estimated and the vocal tract is chiefly one of analogy.

In this paper we attempt to explicitly model the physics of the vocal tract in deriving features for speech recognition. We model the space of valid articulatory configurations using a codebook of articulatory parameters that is derived from electromagnetic articulograph (EMA) measurements of the speech signal. We use the seven-parameter description of the geometry of the vocal tract developed by Maeda [2] to represent each of the articulatory configurations in the codebook.

The set of codewords in the EMA codebook is assumed to be representative of the complete set of valid articulatory configurations for speech. In order to locate the configuration of incoming speech within this space, we use an analysis-by-synthesis approach whereby we generate speech from each of the codewords to best approximate the speech signal. In order to generate speech from the articulatory configuration represented by a given codeword, we use a synthesis technique that explicitly models the physics of the vocal tract. Specifically, we use an articulatory synthesis model developed by Sondhi and Schroeter [3] that uses physiological information of the vocal tract and glottis. We use the spectral error between the synthesized speech from any codeword and the incoming speech signal as indicative of the distance of the articulatory configuration of the signal from that of the codeword. The set of errors from all codewords thus locates the speech signal within the articulatory space. The vector of errors so obtained is then used as an additional feature for speech recognition that complements mel-frequency cepstral coefficients (MFCCs) features, after necessary reductions in dimensionality.

Analysis-by-synthesis approaches have previously been applied to speech recognition. Blackburn [4] used an articulatory codebook that mapped phones generated from N-best lists to articulatory positions. He linearly interpolated the articulatory trajectories to account for coarticulation and used artificial neural networks (ANNs) to map these trajectories into acoustic observations. Each hypothesis was then rescored by comparing the synthesized features to the original acoustic features. Other researchers have attempted to incorporate information about vocal tract processes directly into statistical models that represent speech [5]. We believe that the work described here represents the first attempt to capture explicitly the intrinsic physics of the speech generation mechanism in the feature generation process itself. Experiments in phoneme recognition reveal that significant improvements can be obtained using our articulatory features.

In Section 2 we describe in detail the method we use to generate a codebook representing valid articulatory configurations. In Section 3 we describe the analysis-by-synthesis mechanism we use to derive articulatory features for use in speech recognition. In Section 4 we describe a set of experiments that evaluates different forms of articulatory feature generation within the analysis-by-synthesis formalism. Finally, in Section 5 we present our observations and conclusions.

This work was supported by NSF (Grant IIS-0420866).

2. GENERATING AN ARTICULATORY CODEBOOK

As a first step in our approach, we need to derive a codebook of articulatory configurations that spans the space of all valid configurations for a given set of speakers. To do so, we utilize a set of actual articulatory measurements. We map these measurements to a lower-dimensional representation that we finally cluster into a codebook. We describe each of these steps below.

2.1. Deriving Articulatory Measurements

We obtain a set of realistic articulatory configurations from an actual database of articulatory measurements. In our work we use the MOCHA (MultiCHannel Articulatory) database [6] for this purpose.

The MOCHA database comprises a set of articulatory measurements and corresponding audio recordings from a set of 40 speakers reading 460 TIMIT utterances (British English). The articulatory measurements include electromagnetic articulograph (EMA), electroglottograph (EGG), and electropalatograph (EPG) measurements. For the work reported in this paper, we only use the EMA measurements. The EMA channels include (x, y) coordinates of nine sensors directly attached to the lower and upper lips, lower and upper incisors, tongue tip, tongue body, tongue dorsum, soft palate (velum), and bridge nose. The EMA data is sampled at 500Hz and the corresponding audio is sampled at 16kHz. For our work, we downsample the EMA further to 100 Hz to match the frame rate with which the audio channel is analyzed for recognition.

2.2. Maeda Parameters: Low-dimensional representation of Articulatory Configurations

The Maeda [2] model uses seven parameters to describe the vocal tract shape and compute the areas of the sections of the acoustic tube used in speech generation. Using a factor analysis of 1000 frames of cineradiographic and labiofilm data, Maeda derived a representation of the vocal tract profile as a sum of linear basis vectors or components in a semipolar coordinate space spanning the midsagittal plane of the vocal tract. Each of these components corresponds to the parameters listed in Table 1.

In our work we use Maeda parameters as a seven-dimensional representation of vocal tract configurations. EMA measurements from the MOCHA data are hence converted to these seven-dimensional vectors. To do so, we have developed a geometric mapping from the EMA measurements to Maeda parameters. For P1, we compute the distance between the lower and upper incisors. For P2, we use the horizontal distance between the tongue dorsum and the upper incisor. For P3 we compute the angle between the line joining the tongue tip and the tongue body, and the line joining the tongue body and the tongue dorsum. For P4 we compute the vertical distance between the upper incisor and the tongue tip. For P5 we compute the distance between the upper and lower lips. For P6, we compute the distance between the midpoint of the upper and lower incisors and the line joining the upper and lower lips. Since we are only using the EMA data, we set P7, which pertains to the larynx height, to zero in the rest of the experiments. These parameters are then normalized using their mean and variance, per utterance, to fall within the [-3,+3] range as required by the Maeda model. We use the energy in the audio file to set the starting and ending time of the normalization. This way we exclude the regions where the EMA sensors are off from the steady state position before and after the subject is moving his/her articulators.

Table 1. *Maeda Parameters.*

Parameter	Description	Movement
P1	jaw position	vertical
P2	tongue dorsum position	forward or backward
P3	tongue dorsum shape	roundedness
P4	tongue tip	vertical
P5	lip height	vertical
P6	lip protrusion	horizontal
P7	larynx height	vertical

2.3. Codebook Preparation

Once all measured articulatory configurations are converted to their corresponding Maeda equivalents, we compute a codebook of articulatory parameters. Since P7 is not measured, we do not consider it in this process. The EMA data, and hence the derived Maeda parameters, are aligned with the audio data. To cancel out effects of varying speech rate and phoneme length on the set of available articulatory configurations, we resample the sequence of Maeda parameter vectors to obtain exactly five vectors from each phoneme. To do so, the boundaries of all phonemes in the data must be known. In our work these are obtained by training a speech recognizer (the CMU Sphinx) with the audio component of the MOCHA database, and forced-aligning the data with the trained recognizer to obtain phoneme boundaries.

We sample each phoneme at five positions: the beginning, middle, end, between beginning and middle, and between middle and end, and read the corresponding Maeda parameter vectors. We perform Kmeans over the set of parameter vectors obtained in this manner. We designate the vector closest to the mean of each cluster as codeword representing the cluster. This is done to guarantee that the codeword is a legitimate articulatory configuration. The set of codewords obtained in this manner is expected to span the space of valid articulatory configurations.

3. DERIVING ARTICULATORY FEATURES

Once a codebook spanning the space of valid articulatory configurations is obtained, it is used within an analysis-by-synthesis framework for deriving a feature vector. In the subsections below we describe the synthesis technique employed and how it is used to derive articulatory features.

3.1. Synthesis

The Maeda model converts each vector of articulatory configurations to a vector of areas and lengths of the sections of the acoustic tube describing the shape of the vocal tract. For effective analysis-by-synthesis feature computation, we now need a mathematical model that explicitly implements (or approximates) the physical equations that describe the speech generation mechanism. In our work we then apply the Sondhi and Schroeter [3] model[1] which uses the chain matrices approach to model the overall transfer function of the vocal tract. The transfer function of each section is modeled by a matrix whose coefficients depend on the area and length of the section and on the losses parameters. The input (and output) of the matrix is the pressure and volume velocity in the frequency domain. The transfer

[1] We use the implementations of the Maeda model and the Sondhi and Schroeter model provided with the articulatory synthesis package developed by Riegelsberger [7].

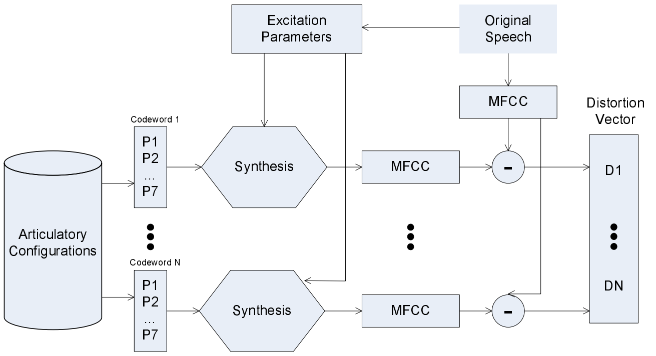

Fig. 1. *Articulatory features computation framework showing two codewords only.*

function resembles the wave equation at each section. The overall transfer function is the product of the matrices.

The glottal source and interaction with the vocal tract is modeled in the time domain using the two-mass model of vocal cords developed by Ishizaka and Flanagan [8]. The parameters of this model are the lung pressure Ps, the glottal area A0, and the pitch factor Q. The overall transfer function must be excited in order to generate speech. In Section 4 we describe results obtained with a variety of different excitation models.

The Sondhi and Schroeter model also allows for the nasal tract coupling to the vocal tract by adjusting the velum opening area. In this paper we assume that the velum is closed since the EMA measurements provided by the sensor attached to the velum are missing or corrupted for many speakers.

3.2. Computing a Feature Vector

For each incoming frame of speech, a corresponding frame of speech is generated by the synthesis model for the articulatory configuration defined by each codeword. Thus there are as many frames of speech synthesized as there are codewords in the codebook. Each frame of synthesized speech is compared to the incoming signal to obtain a distortion value. We use the mel-cepstral distortion, as defined in [9], between the incoming and synthesized speech as the distortion metric in this paper.

The set of distortion values (one per codeword) represents the distance of the incoming signal from each of the articulatory configurations in the codebook, and effectively locates the signal in the articulatory space. A vector formed of the distortion values thus forms our basic articulatory feature vector. The process of creating articulatory feature vectors is shown in Figure 1.

The articulatory feature vector obtained in this manner tends to be high-dimensional – it has as many dimensions as codewords. Its dimensionality is then reduced through linear discriminant analysis (LDA). Other linear or non-linear dimensionality reduction mechanisms may also be employed.

4. EXPERIMENTS AND RESULTS

We conducted a number of experiments to evaluate the usefulness of the proposed articulatory feature extraction method for speech recognition. In order to avoid obfuscating our results with the effect of lexical and linguistic constraints that are inherent in a continuous speech recognition system, we evaluate our features on a simple phoneme classification task where the boundaries of phonemes are assumed to be known. All classification experiments are conducted using simple Gaussian mixture classifiers.

We choose as our data set the audio recordings from the MOCHA database itself, since it permits us to run "oracle" experiments where the exact articulatory configurations for any segment of sound are known. Of the 40 speakers recorded in MOCHA, data for only ten has been released. Of the ten, data for three has already been checked for errors. We checked the data from the remaining seven ourselves, and retained nine for our work: "faet0", "falh0", "ffes0", "fjmw0", "fsew0", "maps0", "mjjn0", "msak0", and "ss2404". Five of the speakers are females and four are males. We checked the EMA, the audio, and the corresponding transcript files for the nine speakers. We discarded the utterances that had corrupted or missing EMA channels, corrupted audio file, or wrong transcripts. We ended up with 3659 utterances, each is around 2-4 secs long. We chose to test on the female speaker "fsew0" and the male speaker "maps0" and train on the rest. All experiments are speaker independent. The amount of training utterances is 2750 and testing is 909. Only the training speakers were used to compute the articulatory codebook. The codebook consisted of 1024 codewords.

In all experiments, the audio signal was represented as 13-dimensional MFCC vectors. We trained a mixture density with 64 Gaussians to represent each phoneme. Cepstral mean normalization (CMN) was applied. No first or second order derivatives were used as they were not found to be useful within the GMM framework.

4.1. An Oracle Experiment: Experiment 1

In this experiment we assume that the exact articulatory configuration (expressed as a vector of Maeda parameters) for each frame of speech is known, and simply obtain it directly from the EMA measurement for the frame. The articulatory feature vector for any frame of speech is obtained simply by computing the Mahalanobis distance between the known Maeda parameter vector for the frame and each of the 1024 codewords in the codebook. We reduce the dimensionality of the resultant 1024-dimensional vectors to 20 dimensions using LDA. A mixture of 32 Gaussians is trained to represent the distribution of these 20-dimensional vectors for each phoneme. The phoneme $\hat{C}$ for any segment is estimated as:

$$\hat{C} = \text{argmax}_C P(C)P(MFCC|C)^{\alpha}P(AF|C)^{(1-\alpha)} \qquad (1)$$

where C represents an arbitrary phoneme, and $MFCC$ and AF represent the set of acoustic and articulatory features for the segment respectively. α is a positive number between 0 and 1 that indicates the relative contributions of the two features to classification. We varied the value of α between 0 and 1.0 in steps of 0.05, and chose the value that resulted in the best classification in the form of phoneme error rate (PER). The classification results and the optimal value of α are shown in Table 2.

Table 2. *PER using MFCC, AF based on oracle knowledge of articulatory configurations, and a combination of the two features.*

Features	fsew0	maps0	Both
MFCC	64.2%	68.1%	66.1%
AF	77.5%	85.8%	81.6%
Combination ($\alpha = 0.85$)	55.2%	62.9%	59.0%
Relative Improvement	14.0%	7.7%	10.8%

We note that feature vectors obtained with oracle knowledge of the vocal tract configuration can result in significant improvements in classification performance in combination with MFCCs, although by themselves they are not very effective.

4.2. Synthesis with Fixed Excitation Parameters: Experiment 2

As explained in Sections 3.1 and 3.2, the articulatory feature vector is computed as the vector of mel-cepstral distortions between the speech signal and the signals generated by the Sondhi and Schroeter model of the vocal tract. The latter, in turn, requires the vocal tract to be excited. In this experiment we assume that the excitation to the synthetic vocal tract is fixed; i.e. the synthesis is independent of the incoming speech itself. This may be viewed as a worst-case scenario for computing features by analysis-by-synthesis.

In this experiment we fixed the excitation parameters [Ps, A0, Q] to the values of [7,0.05,0.9] for a voiced excitation and [7,0.15,0.7] for an unvoiced one. Since the synthesis was independent of the incoming signal, two MFCC vectors were generated from each codeword, one from each excitation. Both synthetic MFCCs were compared to the MFCCs of the incoming speech. Since the energy level in the synthesized speech is fixed, $C(0)$ (zeroth cepstral term) was not considered when computing the distortion. Since two distortion values were obtained from each codeword, the final articulatory feature vector has 2048 dimensions, that were reduced to 20 dimensions using LDA.

The rest of the details of the experiment, including the specifics of dimensionality reduction, distributions estimated and likelihood combination were identical to those in Section 4.1. The results of this experiment are summarized in Table 3.

Table 3. *PER with AF computed using two fixed excitation parameters.*

Features	fsew0	maps0	Both
MFCC	64.2%	68.1%	66.1%
AF	65.9%	72.3%	69.1%
Combination ($\alpha = 0.25$)	60.8%	65.5%	63.1%
Relative Improvement	5.3%	3.8%	4.5%

We note that even in this pathological case, the combination of the articulatory features with MFCCs results in a significant improvement in classification, although it is much less than that obtained with oracle knowledge.

4.3. Excitation Derived from Incoming Speech: Experiment 3

Here we actually attempt to mimic the incoming signal using the various codewords, in order to better localize the incoming signal in articulatory space. To do so, we derive the excitation signal parameters [Ps,A0,Q] from the original signal. Ps (lung pressure) is linearly proportional to the RMS energy. A0 is set based on voicing information. Q is linearly proportional to the pitch value. These excitations are then employed to synthesize signals from each of the 1024 articulatory configurations, which are used to derive a 1024-dimensional articulatory feature vector. As before, the dimensionality of this vector is reduced to 20 prior to classification. $C(0)$ was not considered when computing the distortion. All other details of the classification experiment remain the same as in Section 4.1. Table 4 summarizes the results of this experiment.

We observe that in this "fair" test, the articulatory features are effective at improving classification, providing similar improvements as were obtained with oracle knowledge. Not only are the articulatory features by themselves quite informative (as indicated by the PER obtained with them alone), they also appear to carry information not contained in the MFCCs.

Table 4. *PER with AF computed using excitation parameters derived from the incoming speech.*

Features	fsew0	maps0	Both
MFCC	64.2%	68.1%	66.1%
AF	63.2%	73.1%	68.1%
Combination ($\alpha = 0.6$)	56.9%	64.2%	60.5%
Relative Improvement	11.3%	5.7%	8.5%

5. DISCUSSION

Our results indicate that the analysis-by-synthesis features we introduce do carry information that is complimentary to that contained in the MFCCs. The experiments we report use very a simple statistical model, aimed at highlighting the contributions of these features. It is our hope that these improvements will also carry over to fully-featured HMM-based large vocabulary systems as well. We will explore this possibility as future work.

More importantly, the results indicate that articulatory configurations are intrinsic to phoneme identities. The articulatory features are, in effect, *knowledge-based* representations of the speech signal. Our experiments might thus indicate the need for greater emphasis on the combination of physiologically-motivated knowledge based systems within the statistical framework of speech recognition. This argument is further supported by the fact that while such approaches were not considered feasible in the past due to computational considerations, modern computers make the incorporation of even highly computationally intensive physical models of synthesis into the recognition process feasible. In addition, the advancements in today's machine learning paradigms can be exploited to get better estimates of speech production model parameters, such as the Maeda, and the Sondhi and Schroeter model parameters. Future work will incorporate dynamic constraints on the articulatory configurations to avoid considering all the codewords in the codebook for each frame and reduce the computational complexity required.

6. REFERENCES

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Transac. ASSP*, vol. ASSP-28, No. 4, pp. 357–366, August 1980.

[2] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Modellling*. W.J. Hardcastle and A. Marchal (eds.), 1990, pp. 131–149, Kluwer.

[3] M.M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transac. ASSP*, vol. 35, pp. 955–967, July 1987.

[4] C.S. Blackburn, *Articulatory Methods for Speech Production and Recognition*, Ph.D. thesis, University of Cambridge, 1996.

[5] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 121, pp. 723–742, February 2007.

[6] A. Wrench, "A new resource for production modeling in speech technology," in *Workshop on Innovations in speech processing*, Stratford-upon-Avon, UK, 2001.

[7] E. L. Riegelsberger, *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*, Ph.D. thesis, The Ohio State University, 1997.

[8] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J .*, vol. 51, no. 6, pp. 1233–1268, 1972.

[9] A. Toth and A. Black, "Using articulatory position data in voice transformation," in *ISCA SSW6*, Bonn Germany, 2007.