Improvement of Speech Recognition Method Using Speech Production Mechanism

Jianwu Dang[1,2,3], Yosuke Iizuka[1] , Konstantin Markov[2] and Satoshi Nakamura[2]

[1] Japan Advanced Institute of Science and Technology, Ishikawa, Japan

[2] ATR, Kyoto, Japan

[3] ICP, CNRS, Grenoble, France

E-mail: jdang@jaist.ac.jp, {konstantin.markov, satoshi.nakamura}@art.co.jp

ABSTRACT

This study attempts to combine human mechanisms of speech production into automatic speech recognition (ASR) approaches by using articulatory movements as a constraint. A primary experiment was first conducted on a set of articulatory data, where the articulatory data were treated in the HMM in the same way as the acoustic data. Recognition accuracy increased after adding the articulatory data to the HMM directly. It indicated that the articulatory data have some additional information that is benefit to ASR. We then combined the articulatory data as a hidden parameter in the ASR system built on a hybrid HMM/BN model [1]. Experiments were conducted using this model in monophone recognition with individual models for each speaker and with a uniform model for all speakers, respectively. The accuracy obtained from the HMM/BN was higher than that from the standard HMM without the articulatory data. This study showed a way to incorporate the speech production mechanism in ASR system.

1. INTRODUCTION

Speech is generated and perceived by human in speech communication activities, where human performs almost perfect behaviors in speech processing. However, the main flow of ASR approaches has not taken the human mechanism into account yet. The primary tools used in ASR are the Hidden Markov Models (HMMs) -- they are used to estimate the probability of an acoustic sequence given the model parameters [2]. A nice feature of HMMs is that maximum likelihood techniques provide a way to automatically determine the model parameters from training data. While HMMs have been useful, it has been noted that "[the HMM] is a very inaccurate model of the speech production process" [3].

To account for coarticulations, the common phenomena of speech production, in ASR, a number of models have been proposed as hidden dynamic models [4-6]. Such models describe the physical process of speech production, and attempts to account for the coarticulations and transitions between neighboring frames and phones. Li considered the effects of articulatory movements on speech by modeling the dynamic properties using a quadratic motion equation, and applied this idea in speech recognition [4]. Hogden *et al*. proposed a method so-called MO-MALCOM, that treated the articulation as continuous movements in a virtual speech production space, and used the continuity of the articulation to compensate some discontinuities of acoustic parameters [5]. Gao *et al*. tried to build a uniform model for both speech production and speech recognition via a combination of the Kalman filter and multi-layer perceptron networks [6]. However, these models have not given a satisfactory answer for ASR.

The goal of our study is to incorporate the speech production mechanism in a stochastic model of ASR, so that the automatic parameter estimation can be retained. As the first step, we applied the articulatory data on an ASR system via a combination model of HMM and Bayesian Network [1] to combine the speech production mechanism in ASR. Differing from the studies mentioned above, this study employed observed data of the articulation, which include faithful human mechanisms, instead of the suppositions.

2. A PRIMARY EXPERIMENT ON ARTICULATORY DATA

The articulatory data used in this study were collected using the electromagnetic midsagittal articulographic (EMMA) system at NTT, Japan [7]. Figure 1 shows the placement scheme of the receive coils used in the experiment. Four receive coils were placed on the tongue surface in the midsagittal plane, named T1 through T4, and one coil for each of the upper lip, lower lip, maxilla incisor, mandible incisor (LJ), and the velum, respectively. The sampling rate was 250 Hz for the articulatory channels and 12 kHz for the acoustic channel. The coordinate system is shown in the figure, where the maxilla incisor was chosen as the origin. Speech materials were about 360 Japanese sentences, and three adult male speakers read the sentences at a normal speech rate. The acoustic signal and articulatory data were recorded simultaneously.

To confirm the validity of the articulatory data for the speech recognition purpose, we conducted a primary experiment using both the acoustic data and the articulatory data. The HTK package was employed in the speech recognition with a monophone unit. The first part of the primary experiment was carried out on the articulatory

ISBN 1-876346-48-5

parameters alone. The articulatory data are time-varying vectors with 16 features, which consist of x- and y-coordinates of the eight observation points. To apply the HTK on such data, parameters for the HMM were chosen as 48 dimensions with the displacements, velocities and accelerations of the eight observation points. Features of the articulatory data were written in MFCC format. The second part of the experiment was concerned with the acoustic data alone. The same dimension was applied on acoustic parameters using MFCC with C0 and its first- and second-order coefficients. The results show that the recognition accuracy from the acoustic parameter was higher than that from the articulatory parameters (see Table 1). It implies that the articulatory data have less useful information for speech recognition with the stochastic approaches than acoustic data do.

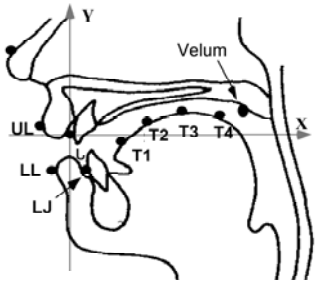

Figure 1. The placement of the reserve coils in the EMA experiment, and the coordinate system used in this study.

However, the articulatory data possibly possess some additional information that is benefit to speech recognition. To examine this conjecture, we constructed the third part of the primary experiment, in which MFCC consists of C0 and its first-order coefficients of the acoustic parameters (32 dimensions), and the 16-dimensional displacement of the articulatory movements. The recognition accuracies are shown in Table 1 for these three situations. The mixture number of the Gaussian distribution in the models ranged from 3 to 16. After the second-order acoustic coefficient was replaced by the articulatory displacement, the recognition accuracy increased more than 2% comparing with the condition of acoustic data alone. This finding indicates that articulatory data possess some beneficial information, which is not covered by acoustic data. Since the articulatory data is not easy to obtain always, the remaining question is how to utilize the limited data in ASR, in other words, how to use the articulatory parameter as a hidden parameter in ASR.

Table 1. The recognition accuracy obtained under three conditions: articulatory data alone, acoustic data alone, and combination of both.

Mixture	Artic. Data	Acost. Data	Acoust.+Artic. Data
3	74.70	80.77	83.92
4	75.01	81.34	84.12
5	75.80	81.81	84.76
8	76.28	82.53	85.64
12	78.42	84.04	86.82
16	79.09	81.93	84.68

3. THE HYBRID HMM/BAYESIAN NETWORK MODEL

Since the articulatory movement is not so easy to obtain as speech sounds. For most of the cases, the articulatory movement can be considered as a hidden parameter. To combine such articulatory features into in a speech recognition model with excellent learning capabilities, the hybrid HMM/BN model proposed by Markov and Nakamura [1] comes in our mind. The hybrid HMM/BN model is a combination of the Hidden Markov Model (HMM) and the Bayesian Networks (BN). In this section, we briefly introduce the hybrid HMM/BN model (see [1] for the details).

In the hybrid HMM/BN model, the temporal characteristics of speech signal are modeled by HMM state transitions, while HMM state probability density is modeled by the Bayesian Network. A configuration of the HMM/BN model is shown in Figure 2.

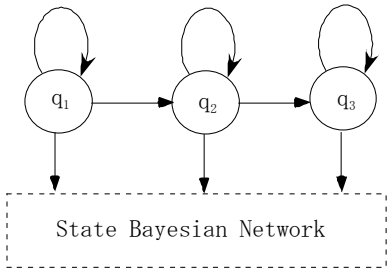

Figure 2. A configuration of the hybrid HMM/BN model

This model is described by two sets of probabilities: HMM transition probabilities $P(q_j|q_i)$ and joint probability distribution of the Bayesian Network $P(X_1,...,X_k)$, where X_i, $i=1,...,K$ are the BN variables. The BN joint probability density function (PDF) can be factorized as:

$$P(X_1,...,X_k) = \prod_{i=1}^{K} P(X_i \mid Pa(X_i)) \qquad (1)$$

where $Pa(X_i)$ denotes the parents of the variable X_i.

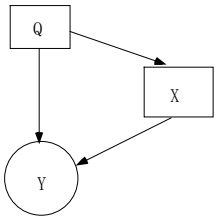

Figure 3. Possible state structure for BN

Figure 3 shows an example of possible state BN structures, where the circle denotes a continuous variable, and the squares are for the discrete ones. BN variables can be either discrete or continuous variables, and some of them can be hidden. These advantages are used in this study for accounting for the articulatory data. Here, variable Q presents HMM state, and Y can be observation vector of speech spectrum. X is the displacement of the observation points on the speech organs obtained from the EMMA

system. The dependency between variables is denoted by the arcs. The relation of Q and Y is described by a conditional probability function. The hybrid HMM/BN model adopted the same training approach as that used in HMM/NN, which is based on the Viterbi training algorithm.

For recognition, the HMM/BN model employs the standard Viterbi decoding algorithm as the conventional HMMs. During this procedure, it requires to calculate input observation likelihood $P(Y|\ Q)$ for each state $Q=q_{ij}$, where i is the HMM index and j is the state index in the ith HMM. For a simple state BN, $P(Y|Q)$ can be derived analytically using the "brute force" inference method. For a more complex state BN, standard exact inference algorithms such as "junction tree" algorithm are available.

4. MODEL LEARNING WITH ARTICULATORY DATA

In this study, we used a monophone HMM/BN model with 3-state left-to-right topology. The structure of the Bayesian Network described above was employed.

4.1 Combining Articulatory Data with the HMM/BN Model

As shown in Fig. 3, the additional information is defined as a discrete variable in the state of the HMM/BN model. It is necessary to discretize the articulatory data for combining them with the model. To do so, the articulatory data were analyzed using the principal component analysis (PCA), and the first four components were used to represent the data. The total contribution ratio of the components was between 80-90%. A vector quantization (VQ) is carried out after the data dimension was reduced. Thus, a discrete variable X is obtained.

State output probability for the BN of Fig. 3 can be calculated from the joint PDF in a closed form shown as formula (2).

$$P(Y,X,Q) = P(Y \mid X,Q) * P(X \mid Q) * P(Q) \qquad (2)$$

Since Y is a continuous variable, $P(Y|X,Q)$ is modeled by the Gaussian density. $P(X|Q)$ can be represented by a conditional probability table (CPT) since X is discrete. Assuming that the articulatory data are known during the training, all BN parameters are fully observable. Thus, Gaussian parameters can be estimated using the ML algorithm, while CPTs are obtained from sample counting.

Since the articulatory data are usually unknown during recognition, variable X should be treated as a hidden parameter. In this case, state output probability becomes

$$P(Y \mid Q) = \frac{P(Y,Q)}{P(Q)} = \frac{\sum_x P(Y, X=x, Q)}{P(Q)} = \sum_x P(X=x \mid Q) * P(Y \mid X=x, Q) \qquad (3)$$

One can see that this expression is actually equivalent to the conventional mixture of Gaussian expression if simply treating the term of $P(X=x|Q)$ as a weight coefficient, where $P(X=x|Q)$ is the ratio of the sample number of class x to the total sample number. After this treatment, the HMM/BN structure degenerates to the structure of the standard HMMs. Thus, the existing HMM decoders can work with the HMM/BN without any modification. Note that there are some differences in the training processing between the HMM/BN and HMM.

4.2 HMM/BN Model Training

The articulatory data set used in this study consists of about 1080 sentences obtained from three male speakers. For this limited data set, the initialization of the HMM/BN is carried out using the bootstrap HMM trained on acoustic features. After the initialization, the HMM/BN model is trained according to the following steps:

1. Phoneme alignment: to perform Viterbi alignment of the training data. It gives a time-aligned state segmentation. This procedure is to prepare training data for the state Bayesian network.

2. BN training: to cluster the acoustic parameter and the articulatory parameter, X. For each class, the state is represented as the Gaussian distribution, and its weight coefficient is calculated by the ratio of the sample number of class x to the total sample number.

3. HMM transition probabilities training: to implement the standard forward-backward training of the HMM transition probabilities.

4. Convergence check: the processing will go back to Step 1 for iterating if the convergence criterion is not met. The training is terminated when the criterion is met.

5. RESULTS FROM THE HMM/BN MODEL

5.1 HMM/BN Model for Individual Speakers

In this study, we first trained HMM/BN model for each individual speaker. To make a comparison between the results with and without articulatory data, recognition experiments were also conducted using the conventional HMM with the acoustic data alone. The book size used in the VQ was designed between 4 and 128, and was transferred into an equivalent mixture number for the comparison. The equivalent number is approximated by the quotient of the actual mixture number of the Gaussian distribution to the state number that was 87 (27 HMM$\times$3 state) in the conventional HMM model. Figure 4 shows the results for three speakers respectively. The dark bars indicate the results using both acoustic and articulatory data (HMM/BN), and the light bars show the results with acoustic data alone (HMM). The basic tendency of the results is that the accuracy obtained from the HMM/BN is higher than that from HMM. As the mixture number get larger, the accuracy generally increases for both conditions. When mixture becomes 16, the result from HMM suddenly got worse. However, almost no damage was seen in the result of the HMM/BN. The recognition accuracy for

ISBN 1-876346-48-5

Speaker 3 is always lower than that from the others, but it shows the same tendency as the others. This experiment reveals two facts: one is that the speech production mechanism is helpful for ASR; and the other is that the HMM/BN model is capable of combining any additional information in an ASR system via automatic learning.

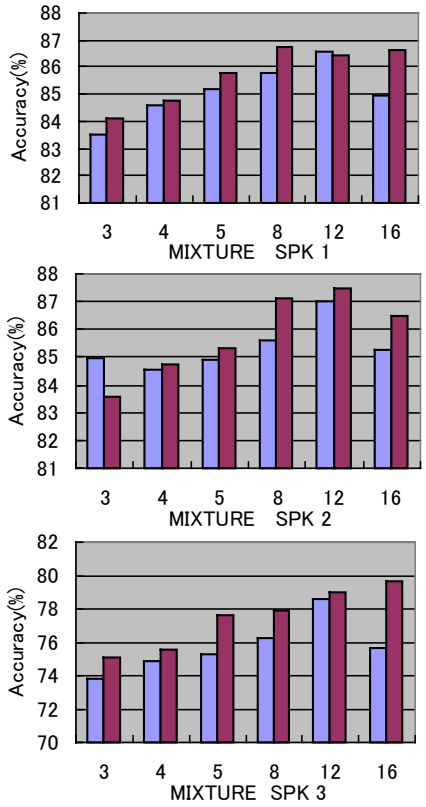

Figure 4. Recognition accuracies using the acoustic data alone (light bars) both acoustic and articulatory data (dark bars) for three speakers

5.2 HMM/BN Model for All Speakers

Learning data for the uniform model were 900 sentences from the three speakers (3×300), and the rest (3×60) served in model testing. Figure 5 shows the results for the uniform model, which are the average values over the three speakers. The line with diamonds denotes the accuracy for HMM with acoustic data alone, and the line with squares shows the results from HMM/BN with both the acoustic and articulatory data. For the uniform model, the HMM/BN model also demonstrated a better performance than the HMM. In this figure, the result obtained in Section 2 is also shown by the line with triangles for a reference. Among these three conditions, the case that the articulatory data replaced partial acoustic parameters in MFCC shows the highest accuracy. This means that the articulatory data still have some potential for increasing the accuracy of ASR.

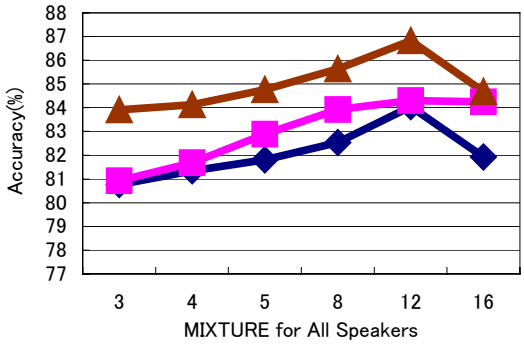

Figure 5. Results obtained from a uniform model. The line with diamonds denotes the results with sound alone, the line with squares for HMM/BN, and the line with triangles for the combination of sound and articulatory data in the same MFCC.

6. DISCUSSION

This study confirmed that articulatory data have some beneficial information to speech recognition, which did not show in speech sounds. The HMM/BN model was employed to combine the articulatory data in speech recognition by an automatic learning, and performed better than the conventional HMM in almost all cases. This study demonstrated a way to apply the speech production mechanism on an ASR system. To promote such a study, a remaining issue is to generate more articulatory data by using a physical articulatory model such as the one proposed by the authors [8].

ACKNOWLEDGMENTS: This research has been supported in part by CREST of Japan Science and Technology. The authors especially thank Masaaki Honda for allowing us to share the articulatory data.

REFERENCES

[1] K. Markov and S. Nakamura. "Large vocabulary ASR system based on the hybrid HMM/BN model," Technical Report IEICE, NLC2002-51, SP2002-128, 43-48, 2002.

[2] F. Jelinek. *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997

[3] K. F. Lee. Automatic Speech Recognition: The Development of the SPHINX System. Boston: Kluwer Academic Publishers, 1989.

[4] L. Deng. "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Communication*. Vol. 24, No. 4, pp. 299-323, 1998.

[5] J. Hogden and P. Valdez. "Bridging the gap between speech production and speech recognition," Proc. of the 5th Seminar on Speech Production, Kloster Seeon, Germany, 2000.

[6] Y. Gao, R. Bakis, J. Huang and B. Xiang. "Multistage coarticulation model combining articulatory formant and cepstral feature," ICSLP-2000, Beijing, China, 2000.

[7] T. Okadome and M. Honda. "Generation of articulatory movements by using a kinematic triphone model," JASA, 453-463, 2001.

[8] J. Dang and K. Honda. "Estimation of vocal tract shapes from speech sounds via a physiological articulatory model," J. Phonetics, 30, 511-532, 2002.