2003－SLP－47　(2)
2003／7／18

Integrating Spectrum and Articulatory features using the hybrid HMM/BN model

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

In this paper, we describe automatic speech recognition system where features extracted from human speech production system in form of articulatory movements data are effectively integrated in the acoustic model for improved recognition performance. The system is based on the hybrid HMM/BN model, which allows for easy integration of different speech features by modeling probabilistic dependencies between them. In addition, features like articulatory movements, which are difficult or impossible to obtain during recognition, can be left hidden, in fact eliminating the need of their extraction. The system was evaluated in phoneme recognition task on small database consisting of three speakers' data in speaker dependent and multi-speaker modes. In both cases, we obtained higher recognition rates compared to conventional, spectrum based HMM system with the same number of parameters.

ハイブリッド HMM/BN モデルに基づいた調音特徴とスペクトル特徴の統合

コンスタンティン　マルコフ[1]、党　建武[1,2,3]、飯塚　陽介[2]、中村　哲[1]

本研究では、我々は音声生成メカニズムにおける調音器官の動きを音声認識システムの音響特徴量に統合し、音声認識性能を改善する方法について提案する。本方法は、ハイブリッド HMM/BN モデルに基づいており、確率的依存関係に基づいて異なる特徴量を容易に統合することができる。さらに、この方法では、認識時に観測できない調音器官の動きの特徴を、観測しないまま隠れ変数として音声認識を行うことができる。本方法を 3 名の発話者からなるデータベースを用いて、特定話者モード、複数話者モードで音声認識実験を行い評価した。その結果、両方の場合において、同一のパラメータ数を用いた場合の従来の HMM による方法に比較して高い性能を得られることを明らかにした。

1. Introduction

Most of the current state-of-the-art speech recognition systems are based on the Hidden Markov Model (HMM) framework where speech is modeled as a sequence of disjoint non-overlapping units. While this approach has been most successful so far it does not take much into account the human speech production mechanism. It has been noted that "[the HMM] is a very inaccurate model of the speech production process" [1].

To account for co-articulations, the common phenomena of speech production, in ASR, a number of models based on hidden dynamic models have been proposed [2-5]. Such models describe the physical process of speech production, and attempt to account for the co-articulations and transitions between neighboring frames and phones. In [2], Deng considered the effects of articulatory movements on speech by modeling the dynamic properties using a quadratic motion equation, and applied the idea in speech recognition. Hogden and Valdez proposed a method called MALCOM [3], that treated the articulation as continuous movements in a virtual

[1] ATR Spoken Language Translation Research Labs
ATR 音声言語コミュニケーション研究所

[2] Japan Advanced Institute of Science and Technology
北陸先端科学技術大学院大学

[3] ICP, Grenoble, France

speech production space, and used the continuity of the articulation to compensate some discontinuities of acoustic parameters. Gao *et al.* [4] 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. In [5], Erler and Freeman proposed HMM based Articulatory Feature Model (AFM) in which each state represents one point in the acticulatory space defined by several hidden articularoty features. Common to all these approaches is that the acticulatory model and features are considered hidden. This allows for eliminating the need of observed articulatory data, which is difficult to collect for training and, in practice, impossible to obtain for recognition.

In contrast, our speech recognition system makes use of observed articulatory data, but only for the acoustic model training. During recognition, the system uses only acoustic data in form of the standard MFCC features. Such scenario is possible when the hybrid HMM/BN model [6] is applied. In this model, articulatory movement data and acoustic speech data are integrated at HMM state level using Bayesian Network (BN), which can model probabilistic dependency between them. The BN parameters are estimated using standard statistical algorithms using both articulatory and acoustic data. In recognition, however, articulatory data are assumed hidden and no observations are required. This allows to fully account for the speech production mechanism in a statistical model, so that the automatic parameter estimation can be retained and a practical system can be built.

Articulatory data

The articulatory data used in this study was 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 coordinate system is shown in the figure, where the maxilla incisor was chosen as origin. The acoustic signal and articulatory data were recorded simultaneously. The sampling rate was 250 Hz for the articulatory channels and 12 kHz for the acoustic channel. The data was collected from three adult male speakers each reading about 360 Japanese sentences at normal speech rate.

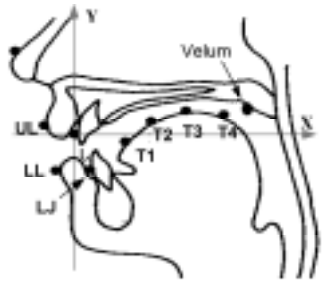

Figure 1: The placement of the reserve coils in the EMA experiment, and the coordinate system used in this study. The gray circles show the observation points in the target vector.

To confirm validity of the articulatory data for the speech recognition task, we conducted a preliminary experiment using the acoustic data and the articulatory data alone as well as both of them together. The articulatory data obtained from the eight observation points are time-varying vectors with 16 components accounting for both the x- and y-coordinates. Thus, our articulatory feature vectors were 48 dimensional (including first and second order coefficients). Using these features we trained 27 monophone 3-state left-to-right HMMs from 900 (3x300) utterances and the rest 180 utterances were used as test data. The same experimental setup was applied with the acoustic data alone. The feature vectors in this case were 16 MFCC coefficients (including C0) and their delta and delta-deltas. In the third case, acoustic and articulatory parameters were combined by replacing the delta-delta coefficients of the acoustic feature vectors with the static articulatory coefficients. Table 1 shows the phoneme recognition accuracy for these three cases using HMMs with different number of mixtures per state.

Table 1: Phoneme recognition accuracy obtained in three cases: acoustic data alone, articulatory 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

The results suggest that articulatory data is not as good as the acoustic data, but when combined together, clear performance gain is observed. This fact indicates that articulatory data posses some additional information which is useful for speech recognition. Since the articulatory data is not easy to be obtained, the remaining question is how to utilize the available limited data.

The hybrid HMM/BN model

One possible answer is to use the hybrid HMM/BN acoustic model [6], which we briefly describe in this section.

This model is essentially a combination of the Hidden Markov Model and Bayesian Network, where the temporal characteristics of speech signal are modeled by HMM state transitions, while HMM state probability density is modeled by the Bayesian Network. The structure of the HMM/BN model is shown in Figure 2.

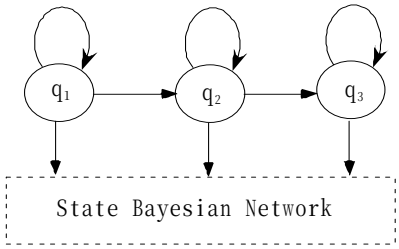

Figure 2: The hybrid HMM/BN model structure.

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 $P(X_1,\ldots,X_k)$, where X_i, $i=1,\ldots,K$ are the BN variables. The BN joint probability density function (PDF) can be factorized as:

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

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

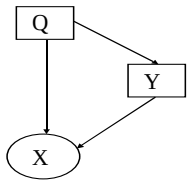

Figure 3: Simple state BN structure. *Q* is the HMM state variable, *X* – speech observation variable and *Y* –some additional variable.

Figure 3 shows an example of a simple state BN structure with three variables. By circle we denote continuous variables, and the squares are used for discrete ones. Therefore, Q and Y are discrete and X is continuous. The arcs represent dependencies between parent and child nodes which can be modeled by Conditional Probability Tables (CPT) if the child is discrete or by Gaussian pdf if the child is continuous.

State output probability for the BN of Fig. 3 can be calculated from the joint PDF in a closed form. According to Eq. 1:

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

If all the BN variables are observable, then state output probability is just $P(X|Y,Q)$ which is one of the BN parameters. However, much more interesting for our task is the case when the additional variable Y is hidden. Then, we are looking for $P(X|Q)$:

$$\begin{aligned} P(X \mid Q) &= \frac{P(X,Q)}{P(Q)} = \frac{\sum_y P(X,Y=y,Q)}{P(Q)} \\ &= \sum_y P(Y=y \mid Q) * P(X \mid Y=y,Q) \end{aligned} \qquad (3)$$

One can see that this expression is actually equivalent to the conventional mixture of Gaussians expression if simply treating the term of $P(Y=y|Q)$ as a mixture weight coefficient for the Gaussian $P(X|Y=y,Q)$. After this treatment, the HMM/BN state output calculation becomes the same as of the standard HMM. Thus, the existing HMM decoders can work with the HMM/BN model without any modification.

Training of the hybrid HMM/BN model is based on the Viterbi algorithm and consists of following steps:

1・**Initialization:** Set initial model parameters randomly or using bootstrap HMM model.
2・**Viterbi alignment:** Obtain time aligned state segmentation of the training data.
3・**BN training:** Train BN using state labeled training data.
4・**HMM transition probabilities updating:** Update HMM transition probabilities using standard forward-backward algorithm.
5・**Convergence check:** Stop, if convergence criterion (training data likelihood increase or preset number of iterations) is met, otherwise go to step 2.

Training of the state BN at step 3 above is done using standard statistical methods. For small networks, when all variables are observable during training, simple ML parameter estimation can be applied. If some of the variables are hidden, then conventional EM algorithm can be used.

2. Integration of the articulatory data

In the previous section we showed how an additional data is used together with the speech observations by employing the hybrid HMM/BN model. Obviously, the articulatory data can be represented by the additional variable Y. The BN of Fig.3, however, requires this additional variable to be discrete. Discretization of the continuous articulatory vectors can be done using standard VQ technique, but at the expense of loosing some resolution accuracy. VQ labels of the articulatory data are, in fact, observations of the additional articulatory variable. Thus, all BN variables are observable for training and the estimation of Gaussian parameters for $P(X|Y,Q)$ can be done through ML algorithm. Weights $P(Y|Q)$ are obtained from label counts. During recognition, articulatory observations are not necessary if HMM/BN state output probability is obtained from Eq. 3 because articulatory variable is hidden.

3. Experiments

In this section, we describe our experimental conditions and report results obtained using speaker dependent and multi-speaker acoustic models.

Common to both cases is the speech and articulatory data processing. Speech front-end was same as the one used in the preliminary experiment we described in Section 2. 12kHz sampled speech wave-forms were framed at 8ms rate with 20ms long Ham-ming window. Feature vector consisted of 16 MFCC coefficients with their delta and delta-deltas. The baseline acoustic only system has 27 phoneme HMMs with 3 states and various number of mixtures and was trained using the HTK toolkit. Since articulatory data were recorded simultaneously and sampled at 250 Hz (which is equivalent to 4 ms frame rate), we used every other articulatory

vector corresponding to one frame of speech. As it is required by our HMM/BN model, articulatory features were quantized using VQ codebooks with sizes ranging from 4 to 128 and trained on the same data. Prior to the vector quantization, articulatory data dimension was reduced from 16 to 4 using PCA analysis technique with a loss of no more than 20% of the information. VQ labels of the quantized articularory vectors were used as training data for the BN articularory variable.

3.1. Speaker dependent model results

From each of the three speakers training data (300 utterances) we trained several HMM/BN models using articulatory VQ codebooks with different sizes. Ideally, the size of the codebook would determine the number of Gaussians per state. However, since the amount of data aligned to different states and having different articulatory labels varies significantly, Gaussians were trained only when this amount of data exceeded empirically set threshold. Thus, different states had different number of mixtures and we use the average mixture per state to describe the model complexity.

Figure 4 shows phoneme recognition results for three speakers for both HMM/BN and baseline HMM models. Although mixture numbers in the figure are integers, actual average mixture number of the HMM/BN is within 10% of those numbers. The test data were the same as those used in the preliminary experiment of Section 2 and consisted of 60 utterances per speaker.

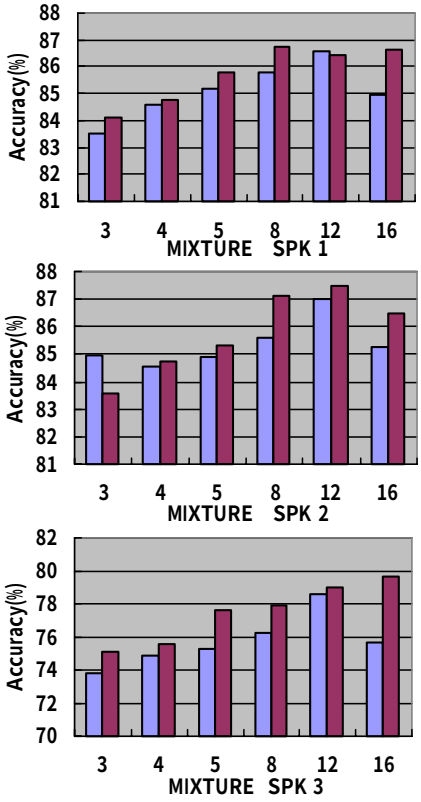

Figure 4: Phoneme recognition accuracies using the acoustic data alone (light-color bars) both acoustic and articulatory data (dark-color bars) for three speakers.

The basic tendency of the results is that the accuracy of HMM/BN is higher than that of HMM. Especially, the accuracy got worse for HMM with 16 mixtures, but there was almost no degradation for the HMM/BN. The recognition accuracy for Speaker 3 is always lower than that from the others, but it shows the same tendency. 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 additional information in an ASR system.

3.2. Multi-speaker model results

One multi-speaker HMM/BN model was trained using the training data from all the speakers (900 utterances). Figure 5 shows the average accuracy over the three speakers for this model (squares) and the baseline acoustic features only HMM (diamonds) in the same phoneme recognition task. Also, shown in this figure is the result obtained in Section 2 using combined acoustic and articulatory feature vectors (triangles). For the multi-speaker case, HMM/BN also performs better than the baseline HMM. However, replacement of partial acoustic parameters by articulatory data in MFCC vectors shows the highest accuracy. This means that there is more potential for utilizing the articulatory data in ASR.

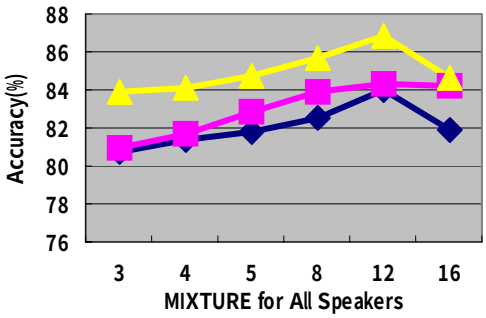

Figure 5: Multi-speaker model results. The diamond line shows baseline HMM results, the square line - HMM/BN, and the triangle line – HMM with combined of acoustic and articulatory feature vectors (from Section 2).

4. Conclusion

This study confirmed that articulatory data have some useful information to speech recognition, which is not included in speech sounds. The HMM/BN model was employed to combine the articulatory data and the experimental results showed its superiority over the conventional HMM in almost all cases. This study demonstrates a way to apply the speech production mechanism in an ASR system.

5. Acknowledgement

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

6. References

1. Lee, K.-F., Automatic Speech Recognition: The Development of the SPHINX System, Kluwer Academic Publishers, Boston, 1989.
2. Deng, Li, “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.
3. Hogden, J. and Valdez, P., “A stochastic articulatory-to-acoustic mapping as a basis for speech recognition”, Proc. IEEE IMTC, Vol.2, pp.1105-1110, 2001.
4. Gao, Y., Bakis, R., Huang, J. and Xiang B., “Multistage co-articulation model combining articulatory, formant and cepstral features”, Proc. ICSLP, pp.25-28, 2000.
5. Erier, K. and Freeman, J., “Using articulatory features for speech recognition”, Proc. IEEE Conference on Commu-nications, Computers and Signal Processing, pp.562-566, 1995.
6. Markov, K. and Nakamura, S., “A Hybrid HMM/BN acoustic model for automatic speech recognition”, IEICE Trans. Inf. & Syst., Vol.E86-D, No.3, pp.438-445, 2003.
7. Okadome, T. and Honda, M., ”Generation of articulatory movements by using a kinematic triphone model”, JASA, pp.453-463, 2001.