Integration of Articulatory Dynamic Parameters in HMM/BN based Speech Recognition System

Konstantin Markov\textsuperscript{1}, Satoshi Nakamura\textsuperscript{1}, Jianwu Dang\textsuperscript{2}

\textsuperscript{1}Spoken Language Translation Research Labs, Advanced Telecommunications Research Institute International, Kyoto, Japan
\textsuperscript{2}Japan Institute of Science and Technology, Ishikawa, Japan
\{konstantin.markov,satoshi.nakamura\}@atr.jp, jdang@jaist.ac.jp

Abstract

In this paper, we describe several approaches to integration of the articulatory dynamic parameters along with articulatory position data into a HMM/BN model based automatic speech recognition system. This work is a continuation of our previous study, where we have successfully combined speech acoustic features in form of MFCC with articulatory position observations. Articulatory dynamic parameters are represented by velocity and acceleration coefficients calculated as first and second derivatives of the articulatory position data. All these features are integrated using the HMM/BN acoustic model where each feature corresponds to different Bayesian Network variable. By changing the BN topology we can change the way articulatory and acoustic parameters are combined. The evaluation experiments showed that the effect of the articulatory dynamic features greatly depends on the BN structure and that careful data analysis is essential in gaining knowledge about the underlying dependencies between different information sources. In comparison with conventional HMM system trained on acoustic data only, the HMM/BN system achieved significant improvement of the recognition performance.

1. Introduction

Most of the current state-of-the-art speech recognition systems are based on speech signal parameterizations which crudely model the behavior of the human auditory system. However, little or no use is usually made of knowledge regarding human speech production system. Research on speech production mechanisms in ASR has been largely focused on using prior phonetic and phonological knowledge and modeling the hidden articulatory trajectories. In many studies, discrete knowledge based features are adopted as articulatory parameterization \cite{1, 2, 3, 4}. They usually describe articulation, e.g. voiced, fricative, nasal, etc. and biomechanics - positions of tongue, lips, jaw and so on. In \cite{1}, such articulatory features are extracted from the parameterized speech signal by means of Neural Networks (NN) and combined with the acoustic features. Knowledge based features can be used to define the HMM state space, as in the Articulatory Feature Model (AFM) \cite{2}. Common disadvantage of such approaches is the quantization of the continuous articulatory parameters where much of the dynamics information is lost. In order to model the co-articulation effect better and to account for the continuous articulatory movement, the discrete articulatory vectors can be regarded as “targets” of trajectory based models. In \cite{3}, Kalman filter is used to smooth target positions and generate “realized” articulations which are further transformed into cepstrum vectors by NN. A stochastic target model is discussed in \cite{5}. In the so called task-dynamic model, articulatory dynamics is described in terms of task-variable which represents vocal tract (VT) construction degrees and locations or VT resonances \cite{6}. Essential issue in building articulatory models with knowledge based features or targets is the selection of the feature set and its size. Too few features may result in very crude and simplistic model. On the other hand, more features will allow for greater precision in trajectory generation, but the complexity of the model and its implementation cost may become prohibitive in practice.

Relatively few studies involve physically recorded articulatory data\cite{7, 8}. Articulatory parameters obtained from actual measurements describe articulation in more fine-grained manner. However, direct observations are usually not available during recognition. Common approach is to estimate articulatory test data from the acoustic signal using NN. In this study, we also make use of actual articulatory data. Rather than trying to learn the mapping between acoustic and articulatory features, we consider them as random variables and model their probabilistic dependency using the hybrid HMM/BN model \cite{9}. In this model, BN represents the states output probability distributions and HMM governs the temporal speech behavior. Articulatory and acoustic parameters are represented by different BN variables. Dependencies are learned from the available training data. During
recognition, however, articulatory variables are assumed hidden which allows for decoding using acoustic observations only. First experiments involving only articulators position parameters were reported previously [10]. Now, we have extended those experiments to include articulatory velocity and acceleration parameters. Various BN topologies integrating those parameters were studied and are described in this paper.

2. The hybrid HMM/BN model

2.1. Brief background

The HMM/BN model is a combination of HMM and Bayesian Network. Speech temporal characteristics are modeled by the HMM state transitions while HMM states probability distributions are represented by the BN. The HMM/BN block diagram is shown in Fig.1. Details about the training and recognition with the HMM/BN model are given in [9, 10].

Figure 1: The HMM/BN model structure. HMM transitions model speech temporal characteristics and BN represents states probability distributions.

2.2. Articulatory Dynamic Parameter Integration

In our previous study [10], we combined speech acoustic and articulatory features using simple BN shown in Fig.2, where variable Q denotes HMM state, X represents continuous MFCC vectors and A is a discrete articulatory variable obtained by vector quantization of the continuous articulatory position data. Now, as articulatory dynamic features we use velocity and acceleration parameters. The most straightforward approach to their integration is to concatenate articulatory position feature vector with velocity and acceleration components, then apply vector quantization and use the same BN as in Fig.2. Assuming that articulatory variable is hidden during recognition, the state output likelihood is calculated as:

\[
p(x_t|q_i) = \sum_{j=1}^{K} P(A = a_j|Q = q_i) \cdot P(X = x_t|A = a_j, Q = q_i)
\]

where \(K\) is the size of the articulatory VQ codebook and \(x_t\) is the acoustic feature vector. If \(P(X = x_t|A = a_j, Q = q_i)\) is Gaussian function, above equation represents mixture of Gaussians where conditional probabilities of the articulatory variable given the state index are the mixture weights. This method, however, does not make use of the BN flexibility and power in modeling data dependencies. We can reasonably assume that MFCC delta coefficients (mostly) depend on articulatory velocity parameters and that MFCC delta-delta coefficients (mostly) depend on articulatory acceleration parameters. A BN which expresses these dependencies is shown in Fig.3 where variables \(X_s, X_v\) and \(X_a\) correspond to MFCC static, delta and delta-delta components. Variables \(A_s, A_v\) and \(A_a\) represent articulatory position, velocity and acceleration parameters. Vector quantization can be done independently for each type of articulatory data using codebooks of different sizes, \(K_s, K_v\) and \(K_a\). According to this BN, acoustic likelihood is calculated as:

\[
p(x_t|q_i) = \prod_{n \in \{s,v,a\}} \sum_{j=1}^{K_n} P(A_n = a^n_j|Q = q_i) \cdot P(X_n = x^n_t|A_n = a^n_j, Q = q_i)
\]

Figure 3: BN structure modeling corresponding dependencies between MFCC static, delta and delta-delta coefficients and articulatory position, velocity and acceleration parameters.

\[
p(x_t|q_i) = \prod_{n \in \{s,v,a\}} \sum_{j=1}^{K_n} P(A_n = a^n_j|Q = q_i) \cdot P(X_n = x^n_t|A_n = a^n_j, Q = q_i)
\]
The output likelihood obtained from this BN is fol-
lish taken into account by making them dependent on each
 articulatory variables. In addition, the possible correlation
features we used 16 MFCCs obtained from 20ms long
data and the rest were left for evaluation. As acoustic
material consist of 350 Japanese sentences read by three
male speakers. 300 of them were selected as training
nt data only. The VQ codebook sizes we chosen such
these experiments are shown in Table 1 along with the
previous results of BN1 trained with articulatory posi-
tion data only. The VQ codebook sizes we chosen such
that all types of models had roughly the same number of
Gaussian mixture components. As the results show, in-
cluding articulatory velocity and acceleration parameters
was effective only with the BN3. The other two showed
degradation of the performance. In the BN1 case where
all three types of articulator y features are concatenated,
the PCA based dimension reduction retains those com-
ponents which have the biggest variance. The data anal-
ysis we did showed that position parameters had lowest
eigenvalues and therefore could be lost in the transforma-

Before the HMM/BN training, all the articulatory
data dimension was reduced to 4 with PCA transforma-
and then they were quantized using VQ codebooks
of sizes ranging from 4 to 1024. VQ labels served as ar-
articulatory observations for the BN training. Initial obser-
vations of the state variable Q were obtained by Viterbi
alignment using the HMM(AC) model. All the HMM/BN
acoustic models have the same structure as the baseline
models except the number mixture components. One it-
eration of BN training was performed and the HMM/BN
transition probabilities were kept the same as in the basel-
line HMM.

First, we evaluated the performance of the HMM/BN
model with BNs of different topologies presented in the
previous section. For convenience, the BN from Fig.2
will be referred to as BN1 and those from Fig.3 and Fig.4,
as BN2 and BN3. To illustrate the effect of articulatory
dynamic features on the model performance, results of
these experiments are shown in Table 1 along with the
previous results of BN1 trained with articulatory posi-
tion data only. The VQ codebook sizes we chosen such
that all types of models had roughly the same number of
Gaussian mixture components. As the results show, in-
cluding articulatory velocity and acceleration parameters
was effective only with the BN3. The other two showed
degradation of the performance. In the BN1 case where
all three types of articulatory features are concatenated,
the PCA based dimension reduction retains those com-
ponents which have the biggest variance. The data anal-
ysis we did showed that position parameters had lowest
eigenvalues and therefore could be lost in the transforma-

<table>
<thead>
<tr>
<th>Speaker 1</th>
<th>Position data only</th>
<th>Position, Velocity and Acceleration data</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN1</td>
<td>86.17</td>
<td>85.75</td>
</tr>
<tr>
<td>BN2</td>
<td>86.44</td>
<td>86.20</td>
</tr>
<tr>
<td>BN3</td>
<td>77.69</td>
<td>77.02</td>
</tr>
</tbody>
</table>

Table 1: HMM/BN phoneme recognition accuracy (%) obtained with three different BN structures using different articulatory feature sets and speaker dependent acoustic models.
tion. The reason for the low results of the BN2 is most probably the fact that acoustic feature vector is split into static, delta and delta-delta parts. This, usually, leads to performance degradation which, in this case, may have diminished the gain provided by the articulatory dynamic parameters.

Next, we investigated the performance of the BN3 as a function of the number of model parameters. By varying the VQ codebooks sizes, we obtained several models with different number of mixture components. As a measure for the model complexity we use the average number of Gaussians per state. Phoneme recognition rates for a model trained on data from the three speakers are plotted in Fig. 5 along with the results obtained from the two baseline HMM models. The performance of the HMM/BN3 is always higher than the HMM(AC), but still not better than HMM(AC+ART). We have to note that HMM(AC+ART) model is of no practical use, because it requires articulatory observations during recognition. Nevertheless, we regard its results as kind of an upper bound for the HMM/BN performance. The plot also shows that the baseline recognition rates start degrading after mixture component number reaches 12 Gaussians per state. In contrast, the best HMM/BN3 results were obtained at roughly two to three times more model parameters. The probable reason is that the baseline HMMs have the same mixture number for each state and given the limited amount of training data, this soon leads to parameter over-training. In the case of HMM/BN3, however, there is a better balance between the amount of training data per state and the number of Gaussians it has, so the over-training appears at bigger number of mixtures.

4. Conclusion

In this study, we investigated several ways of articulatory dynamic features integration in HMM/BN model based speech recognition system. Previously, we successfully combined the MFCC speech features with articulatory position parameters using the same model and the next step in this direction was to expand the system to include the articulatory velocity and acceleration coefficients. The challenge was to find such BN topology that best represents underlying dependencies between the different speech features taking into account their specific characteristics. Evaluation experiments showed that articulatory dynamic parameters can improve ASR performance if integrated properly. Overall, our system achieved much better recognition rates compared to traditional HMM system trained on acoustic data only.

5. Acknowledgment

The research reported here was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan. The authors especially thank Dr. M. Honda for allowing us to share the articulatory data.

6. References