Speech Recognition Based on a Combination of Traditional Speech Features with Articulatory Information

Xugang Lu, Jianwu Dang

School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan 923-1292

{xugang, jdang}@jaist.ac.jp

Abstract

In this paper, we evaluated contributions of the articulatory movement for speech recognition and combined the articulatory information with acoustical features in speech recognition**.** Articulatory movements were observed during utterance of speech sentences, and the speech signals are recorded simultaneously. First, we conducted some speech recognition experiments by combining the acoustic features with articulatory features directly (cascade combination), the performance of the combined features based speech recognition was better than that of the traditional MFCC feature. That implies that articulatory information can give additional information for speech recognition which is not encoded in acoustic features. Second, a combination method was used to combine the articulatory features with traditional speech recognition HMM frame. A Bayesian Network (BN) was added to each state of HMM, articulatory information was represented by the BN as a factor of observed signals. The articulatory information is used during the training of HMM, and then is marginalized as a hidden variable in recognition stage. , Results based on this HMM/BN framework showed that the performance of the method is better than that of traditional speech recognition method.

1. Introduction

Traditional HMM based speech recognition technology has got a lot of progress in these decades, but its performance is still far from real application as what human being can do. Many ideas are proposed to improve the performance, among them, the most possible and reasonable idea is to combine human speech production and perception mechanisms into the recognition system. Human communication system can be regarded as an encoding-decoding system [1],[2], speech production can be regarded as an encoding system while hearing perception can be regarded as a decoding system. Traditional HMM based recognition system is based on statistical pattern recognition technologies. Speech features are extracted, and HMM is trained to get model parameters. During recognition stage, incoming features are compared with the trained model, and the best matched models are chosen as the output. Many human mechanisms are ignored, such as the constraints of speech production system, the properties of hearing perception system, etc. How to use human like mechanisms for speech recognition is the purpose of this research. In this paper, we focus on the application of speech production mechanisms. HMM for speech recognition is a kind of "blind" model, no matter what the incoming stimulations are, all signal units are modeled with uniform models with some hidden states. Some prior information of the signals is ignored, such as the model does not care about how the signals are produced, no matter whether it is speech signal, or acoustic signal of some physical vibrations. Speech signals are a kind of special signals which are produced by human speech organs, if the structures or properties of speech organs are combined into the model as prior information, the model will be refined and be more accurate. Now, speech production mechanism is discussed first.

In Fig.1, the speech production processing is described. High level meaning can be represented by the movement of articulators of production system. In traditional HMM model, only the observed signal level information is used. All information from other levels is ignored.

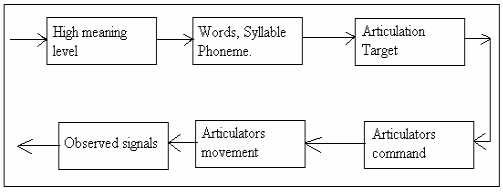

Fig.1 Simplified speech production processing in each level

Since the observed speech signals are produced by the movement of articulators, the mechanical constraints of the articulators must have great effects on the physical properties of the observed signals, that is to say the features of the observed signals will be changed according to different movement patterns of articulators.

The articulators are composed of many parts, such as lip, jaw, teeth, tongue, velum, oral cavity, etc [3]. Different movement patterns of articulators will give different productions of speech signals. Different articulators' movement play different important roles for

producing different kind of sounds, such as for vowels, the tongue shape and positions in the oral cavity do not form a major constriction of air flow, and the variations of tongue placement give each vowel its distinct character[7].In this following part, we investigate further about the contributions of positions of different articulatory movement pattern for speech recognition task.

2. The articulatory data

Articulators' movements are recorded during the production of speech sentences, and the speech signals are recorded at the same time. The recording points are indicated as in Fig.2 using EMMA (Electromagnetic Midsagittal Articulographic).

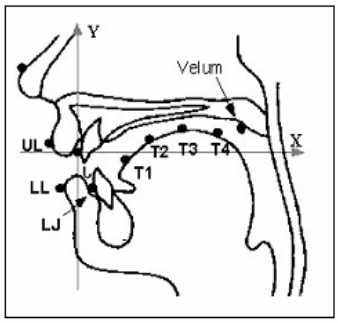

Fig.2 Placement of the sensors used in obtaining the articulatory data.

Each point is recorded with a position coordinate as (x,y), the articulators are: UL for Upper Lip; LL for Low Lip; LJ for Low Jaw; T1 for Tongue point1; T2 for Tongue point2; T3 for Tongue point3; T4 for Tongue point4; Vm for Velum. The sampling frequency is 16k Hz for speech signal data, 250Hz for articulatory data. In Fig.3, an example of the sampling of Japanese sentence is given. We can see from Fig.3, the tongue movement is changing with the different uttering of speech signals. From Fig.3, we can see that the articulator's movement profile has close correlation with the acoustic phoneme unit. In the following, we want to examine the contribution of the movement of articulators to the speech recognition.

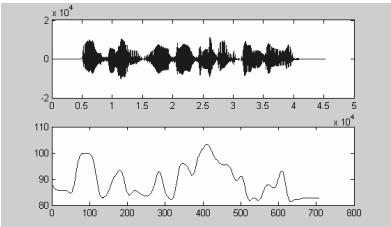

Fig.3 A recorded sentence, upper panel shows speech signal, lower panel is x component of T1

3. Contribution of each observation point

The first experiment is conducted to evaluate the contribution of each observation point to speech recognition. One mixture HMM is used. The feature vector consists of the placement, velocity and acceleration of one observation point alone. Accordingly, the vector with six elements is as

$$[x, y, \Delta x, \Delta y, \Delta\Delta x, \Delta\Delta y]$$

The experimental result is shown in Fig.4

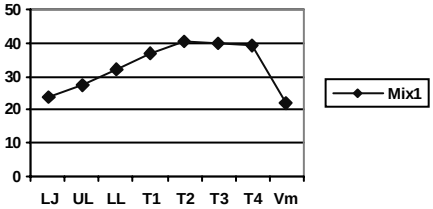

Fig.4 Recognition rate for each observation point. Horizontal-axis is observation point, and vertical-axis is recognition rate(%)

It is obvious that different articulators contribute different information in recognition performance. The tongue contributes more in recognition performance (the experiments can be done in phoneme dependent and phoneme independent conditions). Actually, different articulators play different roles in the production of speech signals. Also, in different recognition task, discriminative features will be represented by different articulators' position. So, in later investigation for how to combine articulatory information into speech recognition, the information from those who contribute more will be taken into consideration. This articulatory information can be combined into HMM model with a combination method.

4. Additional information from articulators and the cascade combination

For further investigation of the information that articulators may provide, we combine traditional speech features (MFCC) with the articulation features together to form new feature for speech recognition. HMM mono-phone recognition system is used. Training database is about 50*6+4 Japanese sentences, another 50 sentences is used for testing. All sentences are read by a Japanese male. About 29 mono-phones HMM model is used including a silence model which is inserted in the start and the end of each sentence. Each phoneme is modeled with 3 left-right HMM model. Mixtures for each state are chosen as 1,2,5,8,12 for each experiment. Three experiments are done:

1. For comparison, MFCC_D_A is used as traditional feature. Feature dimension is 16*3=48.

2. Articulation feature XY_D_A is used, feature vector consists of placement with first and second order derivative of the eight points. The feature vector dimension is also 8*2*3=48.
3. Combination of MFCC (with first order derivative is used) and XY features using MFCC_D+XY, the feature vector dimension is 16*2+8*2=48.

Results are shown in Fig.5. It is clear that articulatory feature has less information than acoustic feature does. So it is not suitable to only use articulatory feature to HMM for speech recognition. But when it is combined with acoustic data, the recognition rate is improved. That is to say, articulation data has some information that acoustic data feature does not have. This information can improve the classification ability of recognition system. In this experiment, articulatory data is combined directly in feature vectors. Actually, articulatory information is another kind of representation for speech signals, it should be used in an additional model to provide extra classification information besides using HMM for acoustic signals.

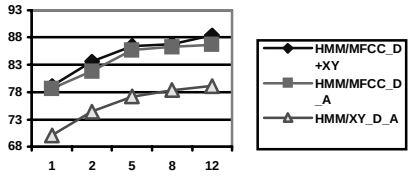

Fig.5 Recognition results for different features. Horizontal-axis is mixture number, vertical-axis is recognition rate(%)

In the above combination for articulatory feature, only articulators' displacement information is used, in the production of speech signals, the dynamic or the movement of the articulators should give further speech discrimination information during the utterance of speech signals. We conduct another experiment to combine MFCC_D with dynamic articulation feature. Instead of choosing only one articulatory channel in our first experiment, we choose three articulatory channels. From Fig.4, we choose Channel 6, 7, 8 (T1-T3) for the experiments, so the articulation feature can be represented as:

$$[X, \Delta X, \Delta\Delta X]$$

Where

$$X = [x6, y6, x7, y7, x8, y8]$$

$$\Delta X = [\Delta x6, \Delta y6, \Delta x7, \Delta y7, \Delta x8, \Delta y8]$$

$$\Delta\Delta X = [\Delta\Delta x6, \Delta\Delta y6, \Delta\Delta x7, \Delta\Delta y7, \Delta\Delta x8, \Delta\Delta y8]$$

It is obvious that the dimension of the articulatory feature is 18, the dimension of MFCC_D is also chosen as 18*2=36. The recognition result is shown in Fig.6.

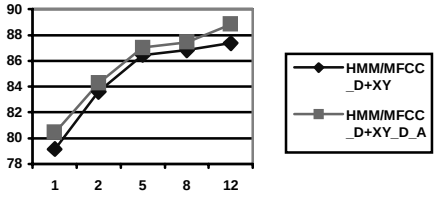

Fig.6 Speech recognition for combining articulators' dynamic information. Horizontal-axis is the mixture number, vertical-axis is recognition rate (%)

From Fig.6, one can see that the recognition results of the new features are better than those features only combining articulation displacement information. So, the velocity and acceleration of the articulators contribute additional information for speech production and play another important role in speech discrimination.

However, the articulatory data is not easy to obtain under a normal condition. So, we must design a combination model in which the articulatory information is used directly for training model parameters, and is used as a hidden parameter during recognition stage. In the next part, this combination model is discussed.

4. Refining HMM to be non-uniform Gaussian mixture model

Usually, for HMM based acoustic model, all modeling units are modeled with uniform topological structures, such as the same state numbers, the same mixture numbers for each state. But in real speech signals, different unit has different kind of temporal structure, time duration, spectral distribution, etc. It is no doubt that the uniform modeling method is inadequate for accurately catch the real statistic property of speech signals. Many auxiliary information can be used for redesigning the HMM structure to refine the HMM. In this paper, the articulatory information can be used to refine HMM (non-uniform Gaussian mixture model for each state of HMM). Articulatory information can be combined with traditional acoustic information to refine the HMM. The combination can be done on different levels, such as on word level, on phoneme level, on state level, etc. as shown in Fig.7. In this paper, the combination level is carried out on state level of each phoneme. For modeling a phoneme, in each state, uniform Gaussian mixture numbers are used, that is to say, the clustering number for each state output is assumed as the same. We can relax this assumption by using different Gaussian mixtures for each state as shown in Fig.8

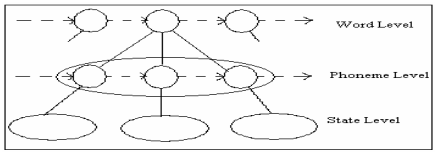

Fig.7 Combination level in Hierachical structure

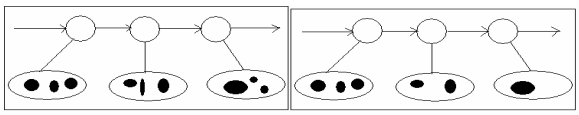

Fig.8 Uniform Gaussian mixture model to non-uniform Gaussian mixture model

How to determine the different clustering numbers for each state? Auxiliary variable or additional information can provide for this kind of clustering. We use the HMM/BN [4],[5][6] model for this purpose. A BN (Bayesian Network) is used as a branch of the state output, and is a condition to produce the observed signals, thus the observation signals can be dependent on two factors: one is the main factor as state variable, another is articulatory information as auxiliary variable which can be used as hidden factor for determining the Gaussian mixture numbers (see [4][5][6] for detail). From the inference algorithm, the articulatory information can be used as an auxiliary variable to control the output clustering for each state. In training HMM, the articulatory information is used to refine the HMM, when in recognition, the articulatory variable can be marginalized as a hidden variable. For convenience of calculation, the articulatory information is represented as a discrete variable, it can be gotten by VQ the articulatory data. Before VQ, a PCA is used to compress the data dimensions from 16 to 4. The experiment is set up as:

- Database: 900 sentences from three speakers (3*300) for training and 3*60 sentences for testing.
- Model: 3 states HMM for each mono-phone, 3-16 Gaussian mixtures for observed outputs.
- Acoustic feature vector is MFCC_D_A with 3*16=48 dimension; Articulatory feature vector is VQ (VQ code size is corresponding to Gaussian Mixture numbers for each experiment)

The experimental result is shown as in Fig. 9. For convenience of comparison, other experimental results are shown in this figure also. Red line is for HMM/MFCC_D+XY_D_A; Black line is for HMM/MFCC_D+XY. Pink line is for HMM/BN/MFCC_D_A&XY(VQ). Blue line is for HMM/MFCC_D_A.

From Fig.9, one can see that the HMM/BN can be used efficiently for incorporating articulatory information and increase the accuracy of speech recognition. Also, from

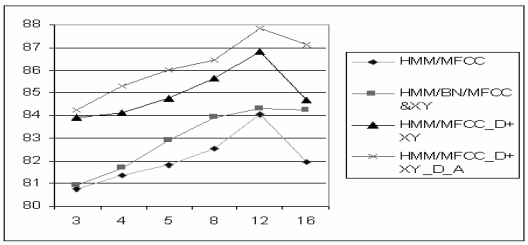

Fig.9 Comparison for different kind of features and combinations. Horizontal-axis is mixture number from 3-16 mixtures. Vertical-axis is recognition accuracy rate(%).

this figure, we can see that there is still a lot of space for improvement to incorporate articulatory information, including the dynamic information of articulators. We have done the experiments by combining the dynamic information of articulators with HMM/BN, but there is no further improvement if the topological structure is not changed. How to incorporate the dynamic information of articulators with HMM for speech recognition is left for our later investigation.

5. Conclusions

Articulators' movement pattern has some classification information that acoustic speech feature does not have. The results from the individual observation points showed that the movement of the tongue body possesses more useful information and contributes more to the speech recognition. Articulatory dynamics can help us to improve speech recognition performance.

6. References

[1] P.Denes and E.Pinson. The speech chain-the physics and biology of spoken language, New York, W.H. Freeman and Company 1993.

[2] Yuqing Gao, et., "Multistage coarticulation model combining articulatory, formant and cepstral features', ICSLP2000, Beijing, China, 2000.

[3] Jianwu Dang, Kiyoshi Honda, "A physical articulatory model for simulating speech production process", Acoustic. Sci. &Tech. 22, 6, 2001.

[4] Konstantin Markov, Satoshi Nakamura, "A hybrid HMM/BN acoustic model for automatic speech recognition", IEICE Trans. Inf.&Syst., Vol. E86-D, No.3, 2003

[5] Jianwu Dang, Konstantin Markov, et. "Application of Articulatory Dynamics on Speech Recognition". Technical Report of IEICE, 2003, ATR, Japan.

[6] Jianwu Dang, Yosuke Lizuka, et. "Improvement of Speech Recognition Method Using Speech Production Mechanism". ICPHS , 2003.

[7] Xuedong Huang, Alex Acero, et. Spoken language processing, Prentice Hall PTR, Upper Saddle River, New Jersey 07458, 2001