NAM-to-Speech Conversion with Gaussian Mixture Models

Tomoki Toda and Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{tomoki, shikano}@is.naist.jp

Abstract

In order to realize a new human communication style using Non-Audible Murmur (NAM) that cannot be heard by people around a speaker, we perform conversion from NAM to ordinary speech (NAM-to-Speech). NAM-to-Speech has a possibility of realizing "non-speech telephone" that is a technique for communicating each other by talking in NAM and hearing in speech. In this paper, we apply a statistical conversion method with Gaussian Mixture Model (GMM) to NAM-to-Speech. In advance, we train GMMs for representing correlations between acoustic features of NAM and those of speech using 50 utterance pairs of NAM and speech. In the conversion, we estimate acoustic spectral and $F_0$ features of speech based on a maximum likelihood criterion, and then synthesize the converted speech with a vocoder. From results of subjective evaluations on intelligibility and naturalness, it is demonstrated that the NAM-to-Speech with GMMs can convert NAM to more consistently natural voice.

1. Introduction

Speech is the ordinary way for most people to communicate. Reasonably, speech is focused on as a man-machine interface. Many researchers have studied speech recognition for several decades and it has dramatically been improved. However, speech has an essential problem as the interface, i.e., it is sensitive to external noise. Many approaches such as speech signal processing and statistical processing have been proposed for improving noise robustness of "speech" recognition.

From another view that the problem of speech recognition is in phonating speech, Nakajima found an indeed different medium from speech, which is Non-Audible Murmur (NAM) [1][2]. Characteristics of NAM are that 1) its power is too small to be heard by anyone around a speaker, 2) it is directly recorded from a skin through the soft tissue with a special microphone such as a stethoscopic microphone, and 3) it can easily be used by anyone who utters speech or used to utter speech [2]. Some results show that NAM recognition, which we call "non-speech recognition", works well by using a specific acoustic model adapted to NAM [3].

We also focus on NAM as a new human communication interface. In recent years, a cellular phone has rapidly spread. It has enabled us to communicate with each other by speech whenever and wherever. However, it has caused a problem. Namely, it is possible that speech is recognized as "noise" by the other people around a speaker in some situations such as a meeting. NAM has a possibility of realizing a new communication style in which we can talk with each other annoying nobody in any situation. An ideal communication form using NAM is that a speaker talks in NAM and a listener hear in ordinary speech, which we call "non-speech telephone." This form is also very effective when talking about private information.

In order to realize it, we need to develop a technique to convert NAM to speech (NAM-to-Speech). This paper describes a method of NAM-to-Speech based on a statistical feature mapping technique. It is well known especially in speech synthesis society that the mapping method based on Gaussian Mixture Model (GMM), which has been proposed by Stylianou [4] as a method of voice conversion, works well. Using a small amount of training data consisting of corresponding pairs of source and target features, this method trains a GMM for modeling correlations between those features in advance. Any sample of the source feature is converted to that of the target feature in a probabilistic way using the trained GMM. It has been shown that the performance of the mapping can be improved by employing maximum likelihood estimation (MLE) considering a dynamic feature [5] and a variance feature [6]. In this paper, we apply the GMM-based mapping with MLE to a mapping from acoustic features of NAM to those of speech used for vocoder-based synthesis. We demonstrate the current performance of the NAM-to-Speech from results of perceptual evaluations on intelligibility and naturalness of the converted speech.

The paper is organized as follows. In Section 2, we describe NAM. In Section 3, the NAM-to-speech based on GMMs is described. In Section 4, experimental evaluations are described. Finally, we summarize this paper in Section 5.

2. Non-Audible Murmur: NAM

NAM is defined as the articulated production of respiratory sound without recourse to vocal-fold vibration, produced by the motions and interactions of speech organs such as the tongue, palate, lips etc., which can be transmitted through only the soft tissue of the head without any obstruction such as bones [2]. It is much easier to understand that NAM is an especial whisper of which power is too small for anyone around a speaker to hear it. NAM is recorded using a stethoscopic microphone attached to the surface of the skin, close behind the ear where is shown in Figure 1. Currently, we use the improved NAM microphone called Open Condenser Wrapped with Soft Silicone (OCWSS) type. That microphone can record frequency components up to 4 kHz while keeping robustness to external noise.
3. Conversion from NAM to Speech: NAM-to-Speech

An acoustic spectrum of NAM is considerably different from that of natural speech. Furthermore, NAM doesn’t have an $F_0$ because it is uttered without the vibration of the vocal codes. Therefore, NAM-to-Speech needs to estimate acoustic features of speech used for synthesis, e.g., spectral envelope, power and $F_0$, from acoustic features of NAM, e.g., spectral envelope and power. Figure 2 shows a schematic diagram of NAM-to-Speech.

We use the same sentence pairs of NAM and speech uttered by a specific speaker to train two GMMs for spectral estimation and $F_0$ estimation. All of the training processes are automatically performed without phonetic information.

3.1. Spectral estimation

We apply the ML-based conversion method considering global variance (GV) of converted parameters [6] to the spectral estimation. We use a spectral feature including power information.

As a source feature of NAM, we use a spectral segment vector $\mathbf{X}$, consisting of spectral feature vectors at several frames around a current frame $t$ [7]. As a target feature of speech, we use a spectral feature vector $\mathbf{Y}_t = [y_{t,0}, \Delta y_{t,0}, y_{t,1}, \Delta y_{t,1}, \ldots, y_{t,\delta}, \Delta y_{t,\delta}]^T$ consisting of static and dynamic features at frame $t$. Using automatically time-aligned source and target features, we train a GMM for representing the joint probability density $p(\mathbf{X}, \mathbf{Y}_t | \Theta)$ [8], where $\Theta$ denotes a set of GMM parameters. We also train another model for representing the probability density of GV of the target static features $p(\mathbf{v}(\mathbf{y})(\Theta_v))$ [6], where $\Theta_v$ denotes a set of parameters of a Gaussian distribution. $\mathbf{y}$ denotes a time sequence of the target static features $[y_{t,1}, y_{t,2}, \ldots, y_{t,\delta}]^T$ over the entire of an utterance, and $\mathbf{v}(\mathbf{y})$ denotes GV, i.e., variances at individual dimensions calculated over the time sequence $\mathbf{y}$.

In the conversion, we estimate the time sequence of the target static feature $\mathbf{y}$ from that of the source feature $\mathbf{X} = [\mathbf{X}_{t,1}, \mathbf{X}_{t,2}, \ldots, \mathbf{X}_{t,\delta}]^T$ so that a likelihood $L = p(\mathbf{Y} | \mathbf{X}, \Theta)|^{-1} p(\mathbf{v}(\mathbf{y})(\Theta_v))$ is maximized, where the constant $\omega$ is a weight. Note that the likelihood is represented as a function of $\mathbf{y}$: the vector $\mathbf{Y} = [Y_{t,1}, Y_{t,2}, \ldots, Y_{t,\delta}]^T$ is represented as $\mathbf{Wy}$, where $\mathbf{W}$ denotes a conversion matrix from the static feature sequence to the static and dynamic feature sequence [9], and the GV vector $\mathbf{v}(\mathbf{y})$ is represented as a quadratic equation of $\mathbf{y}$ [6].

3.2. $F_0$ estimation

We use the ML-based conversion method [5] for the $F_0$ estimation in this paper. We may use GV for the $F_0$ estimation such as the spectral estimation.

We use the same source feature of NAM $\mathbf{X}$, as that used for the spectral conversion. As a target feature of speech, we use static and dynamic features $\mathbf{Y}_t$ of $\mathbf{F}_0$. We train a GMM on the joint probability in a similar way as was described in the spectral estimation.

In the estimation, we determine $\mathbf{y}$ that maximizes a likelihood $p(\mathbf{Y} | \mathbf{X}, \Theta)$ [5]. Again note that this likelihood is represented as a function of $\mathbf{y}$.

4. Experimental Evaluations

In order to demonstrate the effectiveness of the NAM-to-Speech with GMM, we perform experimental evaluations.

4.1. Experimental conditions

We used NAM and speech database uttered by a Japanese male speaker. The 0th through 24th mel-cepstral coefficients [10] were used as a spectral feature at each frame. As a source feature, we used a log-scaled $F_0$ automatically extracted with fixed-point analysis [11]. We set the log $F_0$ value at an unvoiced frame to 0 for convenience sake although such processing was mathematically incorrect. The shift length was 5 ms.

We used 50 utterance pairs of NAM and speech for the training of GMMs. Total durations of NAM and speech are 5.9 minutes (3.1 minutes except for silence frames) and 5.7 minutes (3.6 minutes except for silence frames), respectively.

4.2. Parameter optimization

We optimized parameters such as the number of mixtures in each of GMMs and the number of frames used for constructing the spectral segment feature of NAM. Using 53 utterance pairs not included in the training data, we determined the optimum values of those taking account into both the conversion accuracy of each acoustic feature and the converted speech quality. Consequently, the number of mixtures in the GMM for the spectral estimation was set to 32 and that for the $F_0$ estimation was set to 4. Full covariance matrices were used for both GMMs. The spectral segment feature of NAM was generated as follows: we prepare a vector by concatenating feature vectors at a current frame to 0 for convenience sake although such processing was mathematically incorrect. The shift length was 5 ms.

As a reference, the estimation accuracy of the mel-cepstrum and that of the log-scaled $F_0$ are shown in Table 1 and Table 2, respectively. The ML-based conversion was used for both estimations. The accuracies of voiced and unvoiced decision are also shown in Table 2. RMS error of the log-scaled $F_0$ was 0.242 when the estimated $F_0$ was constantly set to an average value. These results show that the performance of the $F_0$ estimation is awful although that estimation still works slightly better than using the constant $F_0$.

A simple excitation is generated with a pulse train and noise based on the estimated $F_0$. And then, the converted speech is synthesized with the MLSA filter [10] based on the estimated mel-cepstral coefficients. Figure 3 shows an example of acoustic characteristics of NAM, the converted speech with the NAM-to-Speech, and natural speech.
4.3. Perceptual evaluations

We performed perceptual evaluations on intelligibility and naturalness of the converted speech. The following 5 kinds of stimuli were used for the evaluations,

NatSpg-Nat\(F_0\) : synthetic speech with extracted mel-cepstrum and \(F_0\) from natural target speech,
EstSpg-Nat\(F_0\) : synthetic speech with the estimated mel-cepstrum and the extracted \(F_0\),
EstSpg-Est\(F_0\) : synthetic speech with the estimated mel-cepstrum and \(F_0\),
EstSpg-Const\(F_0\) : synthetic speech with the estimated mel-cepstrum and the constant \(F_0\),
NAM : natural NAM.

We assume “NatSpg-Nat\(F_0\)” to be the converted speech with the ideal spectral and \(F_0\) estimations and “EstSpg-Nat\(F_0\)” to be the converted speech with the ideal \(F_0\) estimation. Results of “EstSpg-Est\(F_0\)” show the current performance of the NAM-to-Speech. “EstSpg-Const\(F_0\)” is the converted speech with another NAM-to-Speech method without the \(F_0\) estimation.

All of those were synthesized with duration of NAM. When using the extracted features, we controlled the duration of those based on an automatic time-alignment between the converted mel-cepstrum and the extracted.

In order to reduce burdens of listeners, we used fragments of utterances divided by pauses. Fifty utterance fragments were used, which were not included in the training data.

4.3.1. Evaluation on intelligibility

We performed a perceptual test on the intelligibility by dictation. We allowed listeners to replay the same stimulus time after time. Ten Japanese listeners who have never listened to NAM participated in the test.

Table 3 shows the word correct and the average number of replays by listeners. The word correct of the converted speech with the NAM-to-Speech “EstSpg-Est\(F_0\)” is slightly better than that of NAM “NAM”. Moreover, we can see that the average number of replays of it is smaller than that of NAM. These results show that catching NAM is harder than catching the converted speech. Therefore, the NAM-to-Speech is useful for making the communication smooth.

The accuracy of the spectral estimation is not sufficient because it is observed that the intelligibility in “EstSpg-Nat\(F_0\)” is worse than that in “NatSpg-Nat\(F_0\)”. From a comparison between “EstSpg-Const\(F_0\)” and “EstSpg-Est\(F_0\)”, it is shown that the \(F_0\) estimation causes the improvement of the intelligibility, which might be caused by the effect of the U/V estimation included in the \(F_0\) estimation.
Table 3: Word correct and the average number of replays by listeners

<table>
<thead>
<tr>
<th></th>
<th>Word correct [%]</th>
<th>Number of replays</th>
</tr>
</thead>
<tbody>
<tr>
<td>NatSpg-Nat$F_0$</td>
<td>98.42</td>
<td>1.76</td>
</tr>
<tr>
<td>EstSpg-Nat$F_0$</td>
<td>90.49</td>
<td>3.68</td>
</tr>
<tr>
<td>EstSpg-Est$F_0$</td>
<td>84.29</td>
<td>4.57</td>
</tr>
<tr>
<td>EstSpg-Const$F_0$</td>
<td>80.26</td>
<td>5.53</td>
</tr>
<tr>
<td>NAM</td>
<td>83.86</td>
<td>5.28</td>
</tr>
</tbody>
</table>

4.3.2. Evaluation on naturalness

We performed an opinion test on the degradation of naturalness of the following voices, “EstSpg-Nat$F_0$”, “EstSpg-Est$F_0$”, “EstSpg-Const$F_0$” and “NAM”, compared with the analysis-synthesized voice, “NatSpg-Nat$F_0$”, using a 5-point scale such as 1: very annoying, 2: annoying, 3: slightly annoying, 4: perceptible, but not annoying, and 5: imperceptible. Eight Japanese listeners participated in the test.

A result of the test is shown in Figure 4. The converted speech with the NAM-to-speech “EstSpg-Est$F_0$” is more consistently natural than NAM “NAM”. The scores of NAM widely varies due to the large different results among listeners.

From a comparison between “NAM” and “EstSpg-Const$F_0$”, it is shown that the conversion with the constant $F_0$ causes the large degradation. Therefore, the $F_0$ estimation still works in view of the improvement of naturalness although the estimation accuracy is absolutely insufficient as mentioned above. Because the difference of naturalness between “EstSpg-Nat$F_0$” and “EstSpg-Const$F_0$” is also large, improving the $F_0$ estimation is very important work in NAM-to-Speech but very hard.

5. Conclusions

This paper described a method for converting Non-Audible Murmur (NAM) to ordinary speech, i.e., NAM-to-Speech, with Gaussian Mixture Models (GMMs). NAM-to-Speech is a potential technique to realize “non-speech telephone.” We trained a GMM for the spectral estimation and that for the $F_0$ estimation respectively using a small amount of training data consisting of the same sentence pairs uttered in NAM and speech by a specific speaker. Using both GMMs, we estimated spectra and $F_0$s of speech from spectral segments of NAM based on a maximum likelihood criterion. We performed perceptual evaluations on intelligibility and naturalness of the converted speech. As a result, it was shown that the NAM-to-Speech with GMMs effectively works but its performance is still insufficient. It is inevitable to improve the performance of the NAM-to-Speech. Moreover, we will investigate the conversion from various body-transmitted sounds such as not only NAM but also whisper, murmur and ordinary speech to speech.

Acknowledgment: This research was supported in part by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists and MEXT e-Society leading project. The authors are grateful to Dr. Yoshitaka Nakajima of ATR in Japan for providing NAM and speech database and Prof. Hideki Kawahara of Wakayama University in Japan for permission to use the $F_0$ extraction method.

6. References