INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speaker Dependent Approach for Enhancing a Glossectomy Patient's Speech via GMM-based Voice Conversion

Kei Tanaka[1]*, Sunao Hara*[1]*, Masanobu Abe*[1]*, Masaaki Sato*[2] *and Shogo Minagi*[2]

[1] Graduate School of Natural Science and Technology, Okayama University, Japan
[2] Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan

pjot7wfu@s.okayama-u.ac.jp, {abe, hara}@cs.okayama-u.ac.jp,
sato.masaaki@s.okayama-u.ac.jp, minagi@md.okayama-u.ac.jp

Abstract

In this paper, using GMM-based voice conversion algorithm, we propose to generate speaker-dependent mapping functions to improve the intelligibility of speech uttered by patients with a wide glossectomy. The speaker-dependent approach enables to generate the mapping functions that reconstruct missing spectrum features of speech uttered by a patient without having influences of a speaker's factor. The proposed idea is simple, i.e., to collect speech uttered by a patient before and after the glossectomy, but in practice it is hard to ask patients to utter speech just for developing algorithms. To confirm the performance of the proposed approach, in this paper, in order to simulate glossectomy patients, we fabricated an intraoral appliance which covers lower dental arch and tongue surface to restrain tongue movements. In terms of the Mel-frequency cepstrum (MFC) distance, by applying the voice conversion, the distances were reduced by 25% and 42% for speaker-dependent case and speaker-independent case, respectively. In terms of phoneme intelligibility, dictation tests revealed that speech reconstructed by speaker-dependent approach almost always showed better performance than the original speech uttered by simulated patients, while speaker-independent approach did not.

Index Terms: voice conversion, speech intelligibility, glossectomy

1. Introduction

Speech is the primary means of communication for human beings and plays a crucial role in maintaining one's quality of life (QoL) in everyday life. This is also true for individuals with speech production problems. In this context, intensive studies have been performed to facilitate improvements in the speech of patients with tongue resection or tongue movement disorders. The palatal augmentation prosthesis (PAP) and/or a kinematic artificial tongue (KAT) are ones of such promising methods, and their efficacy have been widely recognized [1,2,3]. Unfortunately, these approaches have the drawback of requiring patients to use a wearable device in their mouth. To solve the problem, we proposed another approach to reconstruct speech quality by using digital signal processing, particularly voice conversion algorithms [4,5,6].

In the previous research [7], we adopted a glossectomy patient as a source speaker and a professional narrator as a target speaker, because professional narrators must utter speech not only with high intelligibility but also with high consistency. Experiment results showed that the proposed algorithm worked well for some phonemes under some phoneme contexts, but did not always work well. One of the reasons for the inconsistent performance might be a speaker factor described as follows. In the voice conversion algorithm, mapping functions between two speakers are trained after finding out correspondences of spectrum features between two speakers using parallel corpus. However, it is difficult to correctly find out the correspondence in our task, because, in the case of speech uttered by a glossectomy patient, spectrum features related to phonemes are quite different from those of normal speech.

To reduce the difficulty, in this paper, we propose to use speech data uttered by the same speaker; i.e., to collect speech uttered by a patient before and after the glossectomy. We call this *speaker-dependent approach*. Because, at the current stage of developing algorithms, it is hard to ask patients to utter the speech in practice, we developed an appliance that restrain tongue movements to simulate glossectomy patients. Normal speakers uttered speech without and with the appliance to simulate speech uttered before and after the glossectomy, respectively. If patients were satisfied with intelligibility of the reconstructed speech, it would be possible to ask them to utter sentences before and after glossectomy.

The rest of the paper is organized as follows. In Section 2, we describe a patient whose speech we focus on reconstructing via voice conversion, and the appliance that restrains tongue movements to simulate glossectomy patients. In Section 3, we explain the conventional voice conversion algorithm [5,6], which is applied to our task. In Section 4, we show our evaluation results and provide a discussion. Finally, in Section 5, we present our conclusions and suggest avenues for future work.

2. Patient, appliance and speech material

2.1. Patient

In April 2014, a 50-year-old man was diagnosed with tongue cancer. After treatment using combination chemotherapy, surgical intervention occurred in June 2014, the surgery involving subtotal glossectomy, right cervical dissection, right cricopharyngeus muscle amputation, and laryngeal elevation. Given a recurrence of the cancer in August 2014, oropharyngeal carcinoma removal surgery, segmental mandibulectomy, mesopharyngeal tumor resection, mandibular bone debridement, and reconstruction with anterolateral thigh flap were undertaken. Nonetheless, the cancer recurred in October 2014, leading to a right

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-841

mandibulectomy, left cervical dissection, reconstruction with free flap of the jawbone rolled letter paper evisceration, and left neck dissection with reconstruction by the right-front outside thigh free flap.

Figures 1 shows an intraoral mock-up following the three operations. The patient was referred to the Department of Oral Rehabilitation and Occlusion, Okayama University for treatment with a PAP. We applied a palatal plate (PP) to the patient's maxilla and a KAT to his mandibular to improve his articulation abilities.

2.2. Appliance

In order to simulate glossectomy patients, we fabricated an intraoral appliance (Figure 2 (a)) which covers lower dental arch and tongue surface to restrain tongue movements during speech (Figure 2 (b)). The appliance was made of pressure thermoforming resin plate. Normal speakers uttered speech without and with the appliance to simulate speech uttered before and after the glossectomy, respectively. To firmly set the appliance, the appliance was developed for each speaker in the same as making a stone model of individual's teeth. Because the appliance blocks for tongue to move above a certain level, normal speakers with the appliance cannot correctly pronounce some phonemes, such as /t/, /d/, /k/, /g/ and etc. This results in imitating glossectomy patients to some extent.

2.3. Speech material

Table 1 and Table 2 show the recorded speech uttered by the patients and four normal speakers (three males, and a female), respectively. The speech recorded in the session 1, 4, and 5 are used to generate mapping functions for voice conversion, and speech recorded in the session 2, 3, 6 and 7 are used for evaluations. Phrase-by-phrase utterance means that speakers read with pauses between phrase, and the reasons why it is used for the training are as follows. Firstly, we would like to reduce burden of patients. The shorter the text, the less likely patients will mispronounce portions of the text. To record correct utterances, repeating sentences several times can place a large burden on patients. Secondary, short utterances can reduce the chance of the failures in finding correspondence by dynamic time warping (DTW). In this paper, we used 100 sentences uttered by phrase-by-phrase for training. On the other hand, for evaluations, we used 50 sentences uttered by sentence-by-sentence. The reason is that, in terms of the number of pauses, sentence-by-sentence utterances are more similar to speech uttered in everyday life than phrase-by-phrase utterances.

2.4. Characteristics of the simulated speech

Figure 3 shows spectrograms of speech uttered by a normal speaker, by the normal speaker with the appliance, and by the patient. As clearly observed in regions marked with red squares, the simulate speech is similar to the patient speech. Mel-cepstrum distances are calculated between the patient and a normal speaker with and without the appliance, after normalizing time duration by DTW. The results are shown in Figure 4. Judging from the results, the Mel-cepstrum distance is decreased when the normal speaker put the appliance on his mouth. The reaming distance might be caused by speaker individuality, i.e., voice quality of male speaker 2 is more similar to the patent than male speaker 1. In conclusion, we can say that the appliance successfully enables for normal speakers to simulate a patient.

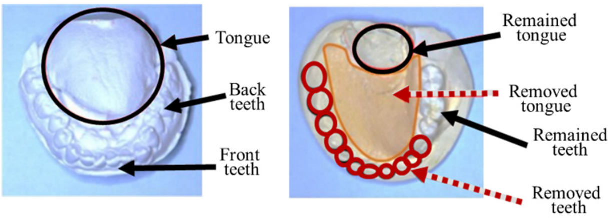
3. GMM-based voice conversion algorithm

3.1. Probability density function

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be D-dimensional source and target feature vectors at frame t, respectively. The joint probability density of the source and target feature vectors is modeled by a GMM as follows:

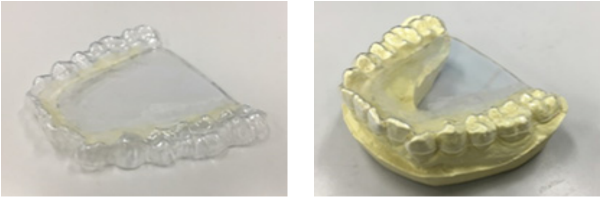
$$P\left(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} w_m N\left(\boldsymbol{z}_t;\, \boldsymbol{\mu}_m^{(z)},\, \boldsymbol{\Sigma}_m^{(z)}\right),$$

where $\boldsymbol{z}_t$ is joint vector $\left[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{y}_t^{\mathrm{T}}\right]^{\mathrm{T}}$, T denotes the transposition of a vector, m is the mixture component index, M is the total number of mixture components, and w_m is the weight of the m^{th} mixture component. Further, the normal distribution with



(a) Before surgeries *(b) After surgeries*

Figure 1: *Intraoral mock-up of a patient*



(a) The appliance *(b) Covered lower dental arch with the appliance*

Figure 2: *The intraoral appliance*

Table 1: *Recorded speech uttered by the patient*

Session	Number of sentence	Speaking style	Devices (PAP and KAT)
1	Full texts (503 sentence)	Phrase-by-phrase utterances	without
2	Subset texts (103 sentence)	Sentence-by-sentence utterances	without
3	Subset texts (103 sentence)	Sentence-by-sentence utterances	with

Table 2: *Recorded speech uttered by normal speakers*

Session	Number of sentence	Speaking style	Appliance
4	Subset texts (103 sentence)	Phrase-by-phrase utterances	with
5	Subset texts (103 sentence)	Phrase-by-phrase utterances	without
6	Sub-subset texts (50 sentence)	Sentence-by-sentence utterances	with
7	Sub-subset texts (50 sentence)	Sentence-by-sentence utterances	without

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $N(\bullet;\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})$. A parameter set of the GMM is $\boldsymbol{\lambda}^{(\boldsymbol{z})}$, which consists of weights, mean vectors, and the covariance matrices for individual mixture components. Joint vectors $\boldsymbol{z}_t$ $(t = 1,2, \dots N)$ are generated by DTW using a parallel speech corpus in which source and target speakers utter the same sentences. Finally, N is the total frame number of training data for the given speech corpus.

Mean vector $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$ of the mth mixture component are written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix},\quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix},$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vectors of the m^{th} mixture component for the source and target, respectively. Matrices $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrices of the m^{th} mixture component for the source and target, respectively. Matrices $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are the cross-covariance matrices of the m^{th} mixture component for the source and target, respectively. The GMM is trained with an expectation-maximization (EM) algorithm using the joint vectors, which are automatically aligned by DTW, in a training set.

3.2. Mapping function

The conditional probability density of $\boldsymbol{y}_t$, given $\boldsymbol{x}_t$, is also represented as a GMM as

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}),$$

where

$$P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{w_m N\left(\boldsymbol{x}_t;\ \boldsymbol{\mu}_m^{(x)},\ \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{n=1}^{M} w_n N\left(\boldsymbol{x}_t;\ \boldsymbol{\mu}_n^{(x)},\ \boldsymbol{\Sigma}_n^{(xx)}\right)}$$

and

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) = w_m N\left(\boldsymbol{y}_t;\ \boldsymbol{E}_{m,t}^{(y)},\ \boldsymbol{D}_m^{(y)}\right).$$

Mean vector $\boldsymbol{E}_{m,t}^{(y)}$ and covariance matrix $\boldsymbol{D}_m^{(y)}$ of the m^{th} conditional probability distribution are written as

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right)$$

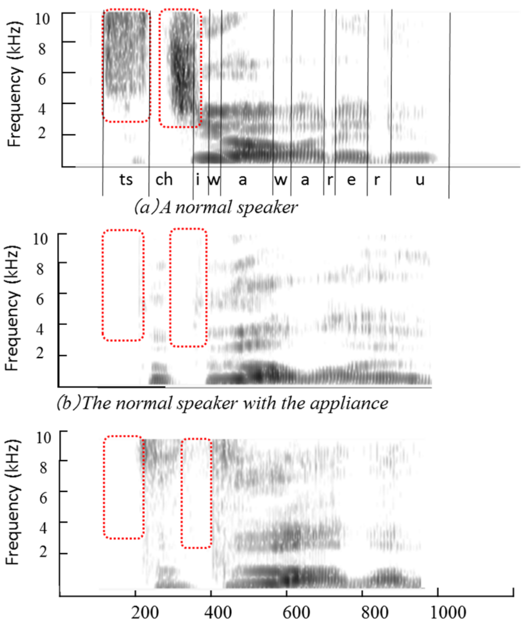
and

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}.$$

Using the conventional method described in [5] and [6], the conversion is performed based on the minimum mean-square error (MMSE) as follows:

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t|\boldsymbol{x}_t] \\ &= \int P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{y}_t\, d\boldsymbol{y}_t \\ &= \int \sum_{m=1}^{M} P(m|\boldsymbol{x}_t,\ \boldsymbol{\lambda}^{(z)})\, P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) \boldsymbol{y}_t\, d\boldsymbol{y}_t \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{x}_t,\ \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)} \end{aligned}$$

Here, $E[\bullet]$ represents the expectation and $\hat{\boldsymbol{y}}_t$ is the converted target feature vector.



(a) A normal speaker

(b) The normal speaker with the appliance

Figure 3: *Spectrograms of speech uttered by a normal speaker, by the normal speaker with the appliance, and by the patient*

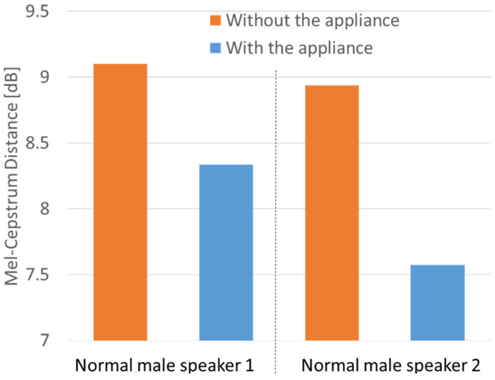


Figure 4: *Mel-Cepstrum distance between the patient and a normal speaker with and without the appliance*

4. Experiments for glossectomy speech reconstruction

Table 3 presents common parameters for the experiments described in this section. The number of mixtures in GMM is based on the results of the previous study [7].

4.1. Experiment 1

We designed this first experiment to ensure the effectiveness of the speaker-dependent approach using objective measures of Mel-frequency cepstrum (MFC) distance. Two kinds of voice conversion were performed, (1) between the male glossectomy patient and a normal male speaker, (2) between a normal male speaker with and without the appliance. As described in Section 2, the mapping functions were trained using phrase-by-phrase speech and evaluated using sentence-by-sentence speech.

The experimental results are shown in Figure 5. In terms of the pre-conversion values, the distance between the glossectomy patient and a normal speaker is much larger than those of the other pairs. This is because the former pair contains both differences caused by glossectomy and speakers, while the other two pairs contains only differences caused by the simulated glossectomy. In terms of after-conversion values, speaker-dependent cases show slightly lower distance than the speaker-independent case. The distances are reduced by 25% and 42% for speaker-dependent cases and speaker-independent case, respectively.

4.2. Experiment 2

To confirm intelligibility improvements by the voice conversion, two sets of dictation tests were performed. The first set was speaker-independent case, i.e., participants were asked to dictate by listening to three kinds of speech: speech uttered by the patient with and without devices (PAP and KAT), and the reconstructed speech by voice conversion. The second set was speaker-dependent case, i.e. the speech presented to the participants were the simulated glossectomy speech and the reconstructed speech by voice conversion. In both sets, fifty phrases were used for each kind of speech and the number of participants was five. For the dictation, participants listened to the speech once, and were allowed to take time as much as they wanted.

Figure 6 and Figure 7 show the mean values and the standard deviations of intelligibility scores for speaker-independent case and speaker-dependent case, respectively. As shown in Figure 6, using PAP and KAT, intelligibilities of fricative, stop, liquid and contracted were improved comparing to those of the original patient speech (pre-conversion). However, original patient speech showed better intelligibility for vowel, Nasal and semivowel. Only for contracted, voice conversion showed better intelligibility than original patient speech. On the other hand, as shown in Figure 7, in the speaker-dependent approach, the after-conversion showed better intelligibility for all phoneme types except liquid. Judging from these results, we can say that speaker-dependent approach works well and influences from different speaker are serious for improving intelligibilities.

5. Conclusions

In this paper, we proposed to generate speaker-dependent mapping functions to improve the intelligibility of speech uttered by patients with a wide glossectomy. Relative to MFC distance of the reconstructed speech, there are a litter differences between speaker-independent approach and speaker-dependent approach. However, in terms of intelligibilities of the reconstructed speech, speaker-dependent approach obviously outperformed speaker-independent approach. This indicates that the speaker factor greatly influences the intelligibility of reconstructed speech.

As part of our future work, we have plans to improve the performance using a segmental approach instead of frame-by-frame approach and using other additional information such as moving image. We would like to also ask patients for cooperation for the developments.

6. Acknowledgements

We express our genuine thanks to Dr. Ken-ichi Kozaki, a professor at the Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan. Regardless of his difficult condition after surgical intervention, he willingly cooperated with us to record his speech. He unfortunately passed away on May 29, 2016. We sincerely pray for the repose of his soul.

Table 3: *Common conditions used in experiments*

Sampling frequency	20 kHz
Speech analysis	STRAIGHT[8]
Frame shift	5 ms
Speech feature	0th~24th mel-cepstral Coefficients and their Δ
The number of mixtures in GMM	16

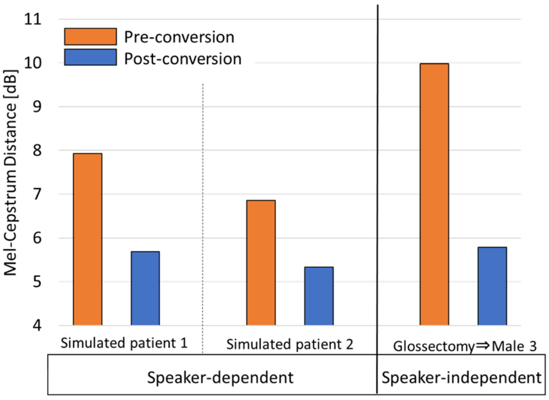


Figure 5: *Mel-Cepstrum distances before and after voice conversion for simulated patients (speaker-dependent) and for the patient (speaker-independent)*

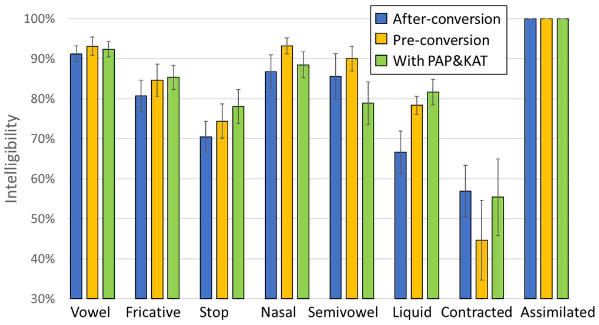


Figure 6: *Intelligibility of the patient (speaker-independent approach)*

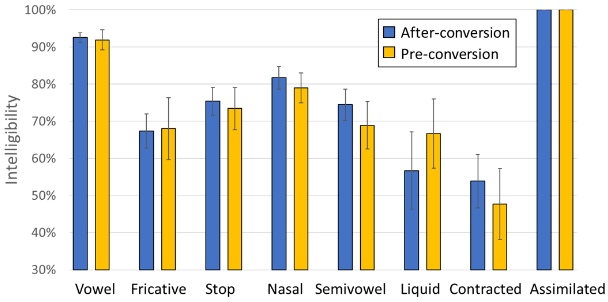


Figure 7: *Intelligibility of a simulated patient (speaker-dependent approach)*

7. References

[1] R. Cantor, T. Curtis, T. Shipp, J. Beume, B. Vogel, "Maxillary speech prostheses for mandibular surgical defects," J. Prosthet Dent, 22:253-60. (1969)

[2] R. Leonard, R. Gillis, "Differential effects of speech prostheses in glossectomized patients," J. Prosthet Dent, 64:701-8. (1990)

[3] K. Kozaki, S. Kawakami, A. Gofuku, M. Abe, S. Minagi et al., "Structure of a new palatal plate and the artificial tongue for articulation disorder in a patient with subtotal glossectomy" Acta Medica Okayama, Vol. 70, No. 3, pp.205-211 (2016.6)

[4] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization," Proc. ICASSP'88, S14.1, pp.655-658. (1988)

[5] Y. Stylianou, O. Capp´e, E. Moulines, "Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Processing," Vol. 6, No. 2, pp. 131–142. (1998)

[6] A. Kain, M. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP'98, pp. 285–288. (1998)

[7] K. Tanaka, S. Hara, M. Abe, S. Minagi , "Enhancing a Glossectomy Patient's Speech via GMM-based Voice Conversion," APSIPA Annual Summit and Conference. (2016)

[8] H. Kawahara, I. Katsue, and A. Cheveigne, "Restructure speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, pp. 187–207. (1999)