A Digital Signal Processor Implementation of Silent/Electrolaryngeal Speech Enhancement based on Real-Time Statistical Voice Conversion

Takuto Moriguchi[1], *Tomoki Toda*[1], *Motoaki Sano*[2], *Hiroshi Sato*[2], *Graham Neubig*[1], *Sakriani Sakti*[1], *Satoshi Nakamura*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Foster Electronic Company, Limited, Japan

{takuto-m, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp, {m_sano, hrssato}@foster.co.jp

Abstract

In this paper, we present a digital signal processor (DSP) implementation of real-time statistical voice conversion (VC) for silent speech enhancement and electrolaryngeal speech enhancement. As a silent speech interface, we focus on non-audible murmur (NAM), which can be used in situations where audible speech is not acceptable. Electrolaryngeal speech is one of the typical types of alaryngeal speech produced by an alternative speaking method for laryngectomees. However, the sound quality of NAM and electrolaryngeal speech suffers from lack of naturalness. VC has proven to be one of the promising approaches to address this problem, and it has been successfully implemented on devices with sufficient computational resources. An implementation on devices that are highly portable but have limited computational resources would greatly contribute to its practical use. In this paper we further implement real-time VC on a DSP. To implement the two speech enhancement systems based on real-time VC, one from NAM to a whispered voice and the other from electrolaryngeal speech to a natural voice, we propose several methods for reducing computational cost while preserving conversion accuracy. We conduct experimental evaluations and show that real-time VC is capable of running on a DSP with little degradation.

Index Terms: statistical voice conversion, real-time processing, reduction of computational cost, DSP, non-audible murmur, electrolaryngeal speech

1. Introduction

Speech communication is one of the most widely used methods for human communication and there is no question that it is a part of our everyday life. However, many barriers still exist in speech communication; e.g., we would have trouble speaking in quiet environments such as in a library as the sound would annoy others; and we may lose the ability to produce a natural voice after undergoing surgery to remove speech organs. In order to break down these barriers, new technologies have been developed, such as silent speech interfaces for allowing people to speak while keeping silent [1, 2, 3, 4] and speaking-aid systems for enhancing unnatural and less intelligible speech produced by vocally handicapped people [5, 6, 7].

Non-Audible Murmur (NAM) [8] has been proposed as one form of the silent speech interface. NAM is a very soft whispered voice, which is acoustically defined as articulated respiratory sounds without vocal-fold vibration conducted through the soft tissues of the head. It is directly detected from the skin surface by attaching a NAM microphone, which is one of the body-conductive microphones, behind the ear. Although NAM can be produced while keeping silent, its sound quality and intelligibility are very low because of the very soft voice and body-conductive recording [9, 10]. To make it possible to use NAM in speech communication, it is essential to make it sound more natural and intelligible.

Electrolaryngeal (EL) speech is produced by an alternative speaking method for laryngectomees whose larynx has been removed by laryngectomy, which is surgery to treat laryngeal cancer. To produce EL speech, the laryngectomee uses an external device called an electrolarynx to mechanically generate excitation sounds. EL speech is quite intelligible but its sound quality is very unnatural owing to the mechanical excitation sounds. Lack of naturalness in EL speech prevents the laryngectomee from smoothly communicating with others. Therefore, it is strongly desired to develop techniques to improve quality of EL speech.

To address these issues, speech enhancement methods based on statistical voice conversion (VC) techniques [11, 12] have been proposed, e.g., silent speech enhancement based on NAM-to-Whisper, which converts NAM into a whispered voice [10], and electrolaryngeal (EL) speech enhancement based on EL-to-Speech, which converts EL speech into normal speech [13]. It has been reported that the trajectory-wise conversion processing [12] is effective for improving naturalness of NAM and EL speech. Moreover, towards the use of these enhancement techniques in human-to-human communication, a low-delay conversion method approximating the trajectory-wise conversion processing with the frame-wise conversion processing has been proposed [14]. Furthermore, a real-time implementation of these enhancement techniques has been proposed and successfully implemented on devices with sufficient computational resources [15]. Towards the practical use of these enhancement techniques, it would be useful to further implement them on devices that are highly portable (e.g., even with no network access) but have limited computational resources.

In this paper, we implement real-time enhancement systems based on VC, such as NAM-to-Whisper and EL-to-Speech, on a digital signal processor (DSP), a highly portable and compact device. Because the computational resources of the DSP are very limited, we propose several methods for reducing the computational cost while preserving conversion accuracy. We experimentally show that the proposed real-time enhancement systems have been successfully implemented on the DSP with little degradation in conversion accuracy.

2. Speech enhancement techniques based on statistical voice conversion

2.1. NAM-to-Whisper and EL-to-Speech

Figure 1 shows the conversion process of NAM-to-Whisper and EL-to-Speech. In NAM-to-Whisper [10], the mel-cepstral seg-

Copyright © 2013 ISCA

ment features of NAM are converted into the mel-cepstrum of a whispered voice. Next, the converted whispered voice is synthesized by filtering white noise excitation signals with the converted mel-cepstrum. As a conversion model for estimating the converted mel-cepstrum of a whispered voice from the mel-cepstral segment of NAM, a Gaussian mixture model (GMM) is used. A parallel data set consisting of NAM and a whispered voice uttered by the same speaker is used to train the GMM.

On the other hand, in EL-to-speech [13], the mel-cepstral segment features of EL speech are converted into not only the mel-cepstrum of normal speech but also F_0 and aperiodic components [16] separately. Next, the converted normal speech is synthesized by filtering mixed excitation signals, which are generated by the converted F_0 and aperiodic components [17], with the converted mel-cepstrum. Therefore, three GMMs are used for estimating the three speech parameters from the mel-cepstral segment of EL speech. To train these GMMs, a parallel data set consisting of EL speech uttered by a laryngectomee and normal speech uttered by a target non-disabled speaker is used.

2.2. Training

To allow conversion to function in real-time, computationally efficient spectral analysis based on the Fast Fourier Transform (FFT) is used to extract the mel-cepstrum of the source speech [15]. Given the mel-cepstral feature vector $\boldsymbol{x}_t$ at frame t, as the source features, a mel-cepstral segment feature vector $\boldsymbol{X}_t$ at frame t is extracted from a joint vector created by concatenating several mel-cepstral feature vectors from frame $t-C$ to frame $t+C$ as follows:

$$\boldsymbol{X}_t = \boldsymbol{E}\left[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top\right]^\top + \boldsymbol{f}, \qquad (1)$$

where $\top$ denotes transposition of the vector. The transformation matrix $\boldsymbol{E}$ and the bias vector $\boldsymbol{f}$ are determined by principal component analysis. On the other hand, to extract target speech parameters, such as mel-cepstrum, log-scaled F_0, and aperiodic components, high-quality speech analysis methods, such as STRAIGHT [18] or mel-generalized cepstral analysis [19], are used because quality of the target speech parameters directly affects quality of the converted speech. Let us assume a feature vector of each target speech parameter $\boldsymbol{y}_t$ at frame t. As the target features, a joint static and dynamic feature vector $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ is created at each frame, where the dynamic feature vector $\Delta\boldsymbol{y}_t$ is calculated as $\boldsymbol{y}_t - \boldsymbol{y}_{t-1}$.

The joint source and target feature vector $\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ is created at each frame by performing time alignment to the parallel data. The joint probability density of the source and target feature vectors is modeled with a GMM as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right), \qquad (2)$$

where $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The parameter set of the GMM $\boldsymbol{\lambda}^{(X,Y)}$ whose total number of mixture components is M is composed of the mixture component weight α_m, the mean vector $\boldsymbol{\mu}_m^{(X,Y)}$, and the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ of each mixture component. At the mth mixture component, the mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ are written as

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \qquad (3)$$

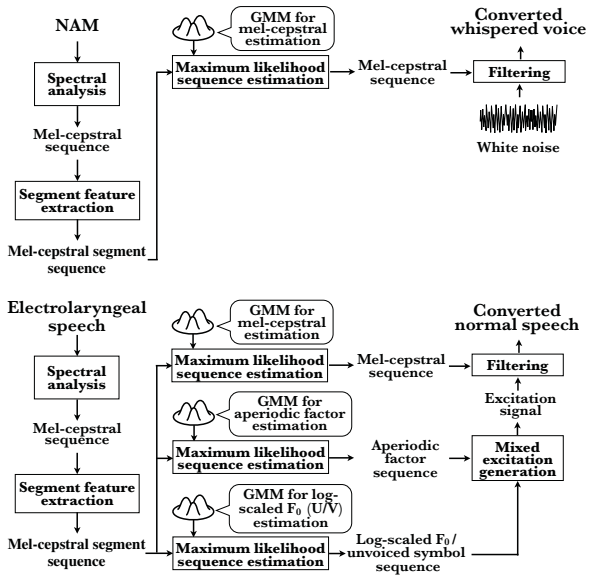

Figure 1: Conversion process of NAM-to-Whisper (upper figure) and EL-to-Speech (lower figure).

where $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$ denote mean vectors of source and target features, respectively. $\boldsymbol{\Sigma}_m^{(XX)}$, $\boldsymbol{\Sigma}_m^{(YY)}$, $\boldsymbol{\Sigma}_m^{(XY)}$, and $\boldsymbol{\Sigma}_m^{(YX)}$ denote covariance or cross covariance matrices of the source and target features, respectively.

2.3. Conversion

Time sequence vectors of the source and target features are denoted as $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$, respectively. The time sequence vector of the converted static features $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is calculated as

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \qquad (4)$$

where $\boldsymbol{W}$ is a transformation matrix that converts a time sequence vector of the static features $\boldsymbol{y}$ into a time sequence vector of the joint static and dynamic features $\boldsymbol{Y}$ [20]. To alleviate muffled speech caused by the over-smoothing effect, global variance (GV) [12] is further considered in conversion into mel-cepstrum of the target speech.

The conversion process given by Eq. (4) is not suitable for real-time processing as it is a batch process using all frames over an utterance. To achieve a real-time conversion process, Eq. (4) is approximated with a low-delay conversion process [14]. First, the suboptimal mixture component m_t at each frame is determined as follows:

$$\begin{aligned} \hat{m}_t &= \underset{m}{\operatorname{argmax}}\, P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) \\ &= \underset{m}{\operatorname{argmax}}\, \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right). \end{aligned} \qquad (5)$$

Next, we determine the converted static feature vector $\hat{\boldsymbol{y}}_{t-L}$ at frame $t-L$ by updating the segment vector of the converted static features $[\hat{\boldsymbol{y}}_{t-L}^\top, \cdots, \hat{\boldsymbol{y}}_t^\top]^\top$ frame by frame using Kalman filtering without considering the GV. Finally, the converted static feature vector is enhanced frame by frame with a GV-based postfilter [15].

3. Reduction of computational cost

3.1. Diagonalization of source covariance matrices

As full covariance matrices are used in NAM-to-Whisper and EL-to-Speech, the computational cost for the mixture component selection given by Eq. (5) is high. Although it still works in real-time on devices with sufficient computational resources, we have found that it does not work in real-time on the DSP. Therefore, we implement diagonalization of the source covariance matrices proposed in [15] to reduce the computational cost. The source covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ are factorized into mixture-dependent diagonal covariance matrices $\boldsymbol{\Sigma}_{m,diag}^{(XX)}$ and a global full transformation matrix $\boldsymbol{A}$ as follows:

$$\boldsymbol{\Sigma}_m^{(XX)} \simeq \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{m,diag}^{(XX)}\boldsymbol{A}^{-\top}. \tag{6}$$

This covariance structure makes the computational cost equal to that necessary when using the diagonal covariance matrices because the global transformation matrix can be applied in advance to the transforms for the feature extraction shown in Eq. (1). As reported in [15], this approximation method tends to cause only a small degradation in conversion accuracy.

3.2. Program optimization

To achieve real-time processing on the DSP, we perform several program optimizations, such as pre-calculation of the twiddle factor of the FFT and the use of DSP-specific functions. These optimizations cause no adverse effect on conversion accuracy. To further reduce the computational cost, some operations are approximated; e.g., exponential and logarithmic functions are approximated with piecewise linear functions; and high-order cepstral coefficients are approximated with zero values to reduce the computational cost for transforming cepstral coefficients to mel-cepstral coefficients using a first-order all-pass filter. These approximations may cause adverse effects on conversion accuracy.

3.3. Increase of frame shift

In real-time conversion processing, all procedures at each frame, such as feature extraction, conversion, and synthesis, should be finished within the frame shift. Although the frame shift is set to 5 ms in the traditional conversion systems [10, 13, 15], we change it to 10 ms. Because feature extraction and conversion are performed only once at each frame, computational costs for these procedures are not changed according to the frame shift. Consequently, the real-time factor calculated as a rate of the processing time divided by the frame shift is reduced by half for these procedures. On the other hand, the computational cost for synthesis increases as 10 ms of the waveform signal needs to be generated at each frame. Moreover, we change a few parameters, such as the number of frames to extract the mel-cepstral segment (related to C in Eq. (1)) and the number of delay frames L in the low-delay conversion described in Section 2.3, to keep latency in conversion almost the same as in 5 ms frame shift. Table 1 shows the latency difference caused by changing frame shift from 5 ms to 10 ms. The frame shift change may cause adverse effects on conversion accuracy due to the decrease of time resolution.

3.4. Simplification of the mixed excitation model

The STRAIGHT mixed excitation model [16] is used in traditional EL-to-Speech systems [13]. It is effective for improving quality of converted speech but it needs to estimate aperiodic components and calculate a weighted sum of a pulse train and white noise excitation with frequency-dependent weighting values determined from the estimated aperiodic components [17]. To reduce the computational cost for generating the excitation signals, a two-band mixed excitation model used in Harmonic plus Noise Model (HNM) [21] is implemented as a simpler mixed excitation model. As shown in Fig. 2, this model generates the excitation signal using a pulse train in low-frequency bands and white noise in high-frequency bands. The maximum voiced frequency is used to define the boundary between these two frequency bands. Although this value normally changes frame by frame, we approximate it with a fixed value (4 kHz in this paper) to further reduce the computational cost. Time-invariant low/high-pass filters designed by a low-order Butterworth filter are used to efficiently generate the mixed excitation signals.

Table 1: Latency difference caused by increasing frame shift.

Frame shift	C	L	$C+L$	Latency [ms]
5 ms	4	3	7	35
10 ms	2	2	4	40

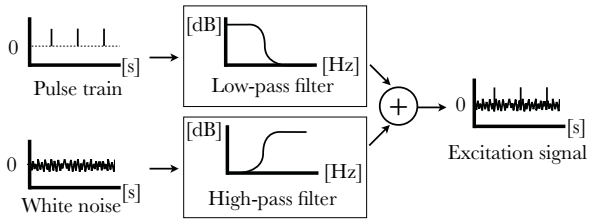

Figure 2: Two-band mixed excitation model.

4. Experimental evaluations

4.1. Experimental conditions

We conducted experimental evaluations of our proposed implementation in NAM-to-Whisper and EL-to-Speech. We used TMS320C6748 (375 MHz) as a floating point DSP.

NAM-to-Whisper: Each of two male speakers and one female speaker recorded 140 newspaper sentences in NAM with NAM microphone and in a whispered voice with an air-conductive microphone. The sampling frequency was 16 kHz. For each speaker, 70 newspaper sentences were used for training and the other 70 newspaper sentences were used for test. The 0th through 24th mel-cepstral coefficients were used as spectral features. The speaker-dependent GMMs for mel-cepstral estimation were trained for the individual speakers. The number of mixture components of each GMM was set to 32. The real-time factor and spectral conversion accuracy with mel-cepstral distortion in the following systems were evaluated:

Offline based on a batch-type process [10],

Baseline based on the conventional real-time implementation without the diagonalization [15],

Diag implementing the diagonalization for **Baseline**,

Fast implementing program optimizations for **Baseline**,

10ms changing frame shift from 5 ms to 10 ms in **Baseline**,

Diag+Fast combining **Diag** and **Fast**,

Diag+Fast+10ms combining **Diag**, **Fast**, and **10ms**.

EL-to-Speech: One laryngectomee recorded 50 phonetically balanced sentences in EL speech and another non-disabled speaker recorded the same sentences in normal voices. The

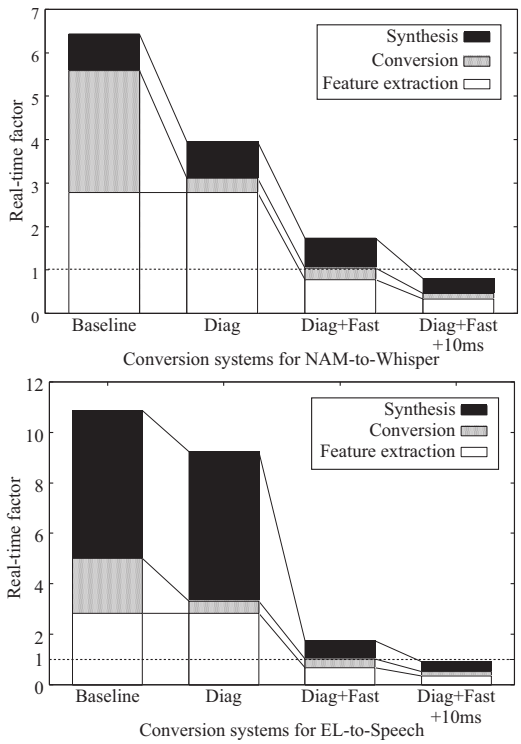

Figure 3: Real-time factor in each system of NAM-to-Whisper (upper figure) and EL-to-Speech (lower figure).

sampling frequency was 16 kHz. Forty sentences were used for training and the other 10 sentences were used for test. The 0th through 24th mel-cepstral coefficients were used as spectral features and log-scaled F_0 and aperiodic components were used as excitation features. The numbers of mixture components of the GMMs were set to 32 for the spectral estimation, 16 for the F_0 estimation, and 16 for the aperiodic estimation. The real-time factor and spectral conversion accuracy with mel-cepstral distortion in **Offline**, **Baseline**, **Diag**, **Fast**, **10ms**, **Diag+Fast**, and **Diag+Fast+10ms** were evaluated. In **Fast**, the simplification of the mixed excitation model as well as the program optimizations was also implemented for **Baseline**. We also conducted an opinion test on naturalness using a 5-scaled opinion score (1: very bad to 5: excellent) as a subjective evaluation. Twelve listeners evaluated naturalness of original EL speech (**EL**) and three types of converted speech by **Offline**, **Baseline**, and **DSP** that is an actual DSP conversion system based on **Diag+Fast+10ms**.

4.2. Experimental results

Figure 3 shows the real-time factor of feature extraction, conversion, and synthesis processing in each system for NAM-to-Whisper and EL-to-Speech. The diagonalization (**Diag**) significantly reduces the real-time factor in conversion. The program optimization (**Diag+Fast**) significantly reduces the real-time factor in feature extraction. In EL-to-Speech, it also significantly reduces the real-time factor in synthesis thanks to the simplified mixed excitation model. The real-time factor in total processing can be successfully reduced below 1 by further changing the frame shift from 5 ms to 10 ms (**Diag+Fast+10ms**). Therefore, **Diag+Fast+10ms** is capable of running on the DSP in real-time.

Figure 4 shows spectral conversion accuracy using the mel-cepstral distortion (MCD) in each system for NAM-to-Whisper

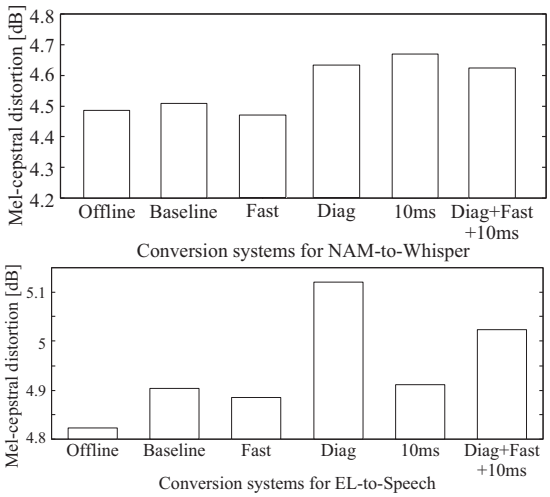

Figure 4: Mel-cepstral distortion in each system of NAM-to-Whisper (upper figure) and EL-to-Speech (lower figure).

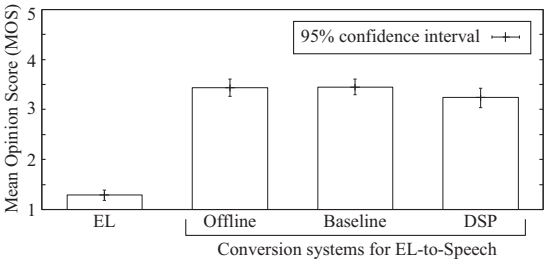

Figure 5: Result of opinion test on naturalness in EL-to-Speech.

and EL-to-Speech. MCD before conversion is 9.28 dB in NAM and 9.24 dB in EL speech. In NAM-to-Whisper, the online conversion system (**Baseline**) does not cause any significant degradation compared with the offline conversion system (**Offline**). Although **Fast** causes no degradation, **Diag**, **10ms**, and **Diag+Fast+10ms** cause around a 0.1–0.2 dB increase in MCD. On the other hand, in EL-to-Speech, **Baseline** causes nearly a 0.1 dB increase in MCD compared with **Offline**. **Diag** and **Diag+Fast+10ms** further cause around a 0.1–0.2 dB increase in MCD. We can see from these results that the real-time conversion systems (**Diag+Fast+10ms**) can be implemented on the DSP while keeping the increase of MCD less than around 0.2 dB and conversion accuracy of the implemented DSP system is still sufficient.

Figure 5 shows a result of the opinion test on naturalness in EL-to-Speech. Although **DSP** causes slight degradation in naturalness compared with **Offline** and **Baseline**, it is small enough to be insignificant in practice. **DSP** still yields significant improvements in naturalness of EL speech (**EL**). We also confirmed that **DSP** does not cause any significant degradation in naturalness compared with **Offline** in NAM-to-Whisper as well although we don't show any results due to space limitation.

5. Conclusions

In this paper, we have proposed several methods for reducing the computational cost of speech enhancement processing based on real-time statistical voice conversion (VC) and have successfully implemented real-time VC systems for enhancing non-audible murmur (NAM) and electrolaryngeal (EL) speech on a Digital Signal Processor (DSP). The experimental results demonstrate that the DSP systems have the capability to significantly enhance NAM and EL speech and run in real-time.

Acknowledement: Part of this work was supported by JSPS KAKENHI Grant Number 22680016.

6. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg, "Silent speech interfaces," *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.

[2] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," *Proc. INTERSPEECH*, pp. 1493–1496, Jeju Island, Korea, Sep. 2004.

[3] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, Vol. 52, No. 4, pp. 341–353, 2010.

[4] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, Vol. 52, No. 4, pp. 288–300, 2010.

[5] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomedical Engineering*, Vol. 57, No. 10, pp. 2448–2458, 2010.

[6] H. Liu, Q. Zhao, M.-X. Wan, and S.-P. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomedical Enginering*, Vol. 53, No. 5, pp. 865–874, 2006.

[7] A.Hisada and H.Sawada, "Real-time clarification of esophageal speech using a comb filter," *Proc. ICDVRAT*, pp. 39–46, Sep. 2002.

[8] Y. Nakajima, H. Kashioka, N. Campbell, K. Shikano, "Non-Audible Murmur (NAM) recognition," *IEICE Trans. Information and Systems*, Vol. J87-D-II, No. 9, pp. 1757–1764, 2004.

[9] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," Speech Communication, Vol. 52, No. 4, pp. 301–313, 2010.

[10] T. Toda, M. Nakagiri, K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 20, No. 9, pp. 2505–2517, 2012.

[11] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

[12] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[13] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, Vol. 54, No. 1, pp. 134–146, 2012.

[14] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[15] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, Portland, USA, Sep. 2012.

[16] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proc. MAVEBA*, Firenze, Italy, Sep. 2001.

[17] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. INTERSPEECH*, pp. 2266–2269, Pittsburgh, USA, Sep. 2006.

[18] H. Kawahara, I. Masuda-Katsue, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a receptive structure sounds," *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[19] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.

[20] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp.1315–1318, Jun. 2000.

[21] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 1, pp 21–29, Jan. 2001.