Exemplar-based Individuality-Preserving Voice Conversion for Articulation Disorders in Noisy Environments

Ryo AIHARA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI

Graduate School of System Informatics, Kobe University, Japan

aihara@me.cs.scitec.kobe-u.ac.jp, takashima@me.cs.scitec.kobe-u.ac.jp,
takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

We present in this paper a noise robust voice conversion (VC) method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movements of such speakers are limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In this paper, exemplar-based spectral conversion using Non-negative Matrix Factorization (NMF) is applied to a voice with an articulation disorder in real noisy environments. In this paper, in order to deal with background noise, an input noisy source signal is decomposed into the clean source exemplars and noise exemplars by NMF. Also, to preserve the speaker's individuality, we use a combined dictionary that was constructed from the source speaker's vowels and target speaker's consonants. The effectiveness of this method was confirmed by comparing its effectiveness with that of a conventional Gaussian Mixture Model (GMM)-based method.

Index Terms: Voice Conversion, NMF, Articulation Disorders, Noise Robustness, Assistive Technologies

1. Introduction

There are 34,000 people with speech impediments associated with an articulation disorder in Japan alone. One of the causes of speech impediments is cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified into the following types: 1) spastic, 2) athetoid, 3) ataxic, 4) atonic, 5) rigid, and a mixture of these types [1].

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers. In the case of a person with this type of articulation disorder, his/her movements are sometimes more unstable than usual. That means, in cases where movements are related to speaking, their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. Veaux et al. used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders [2]. However, because athetoid symptoms also restrict the movement of the sufferer's arms and legs, it is difficult for them to input text information to synthesize their voice. Most people suffering from athetoid cerebral palsy cannot communicate by sign language or writing for the same reason, so there is great need for voice conversion (VC) system for them.

In this paper, we propose a VC method for articulation disorders. Our VC method has the following two benefits. The first benefit is the noise robustness. When we use a VC system in a real environment, background noise is an unavoidable problem. Input noise may degrade the VC performance due to unexpected mapping in the features. Our proposed method includes a noise separation system in VC that enables the VC method to work effectively in noisy environments. The second benefit is to preserve the individuality of the source speakers voice. People with articulation disorders wish to communicate by their own voice if they can. By using an individuality-preserving dictionary, we convert their voice into a well-ordered voice preserving their voice individuality.

A GMM-based approach is widely used for VC because of its flexibility and good performance [3]. This approach has been applied to various tasks, such as speaker conversion [4], emotion conversion [5, 6], speaking assistance [7], and so on. The conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using Minimum Mean-Square Error (MMSE) using a parallel training set. If the person with an articulation disorder is set as a source speaker and a physically unimpaired person is set as a target speaker, an articulation-disordered voice may be converted into a well-ordered voice. However, because the GMM-based approach has been developed mainly for speaker conversion [4], the source speaker's voice individuality is also converted into the target speaker's individuality. Furthermore, the effectiveness of these approaches was confirmed with clean speech data, and the utilization in noisy environments was not considered.

In the research discussed in this paper, we conducted VC for articulation disorders using Non-negative Matrix Factorization (NMF) [8]. NMF is a well-known approach for source separation and speech enhancement. In these approaches, the observed signal is represented by a linear combination of a small number of elementary vectors, referred to as the basis, and its weights. In some approaches for source separation, the bases are grouped for each source, and the mixed signals are expressed with a sparse representation of these bases. By using only the weights of the bases related to the target signal, the target signal can be reconstructed. Gemmeke et al. proposes an exemplar-based method for noise robust speech recognition [9]. In that method, the observed speech is decomposed into the speech bases, noise bases, and their weights. Then the weights of the speech bases are used as phonetic scores instead of the likelihoods of Hidden Markov Model for speech recognition.

In our study, we adopt the supervised NMF approach [10], with a focus on VC from poorly articulated noisy speech resulting from articulation disorders into well-ordered clean articulation. The parallel exemplars (called the 'dictionary' in this pa-

per), which consist of a articulation-disordered exemplars and a well-ordered exemplars, are extracted from the parallel data. Also, the noise exemplars are extracted from the before and after utterance sections in an observed signal. An input noisy spectrum with an articulation disorder is represented by a linear combination of clean articulation-disordered exemplars and noise exemplars using NMF. Only the weights (called 'activity' in this paper) related to the clean exemplars are picked up, and by replacing an articulation-disordered basis with a well-ordered basis, the original speech spectrum is replaced with a well-ordered spectrum. Moreover, in the voice of a person with an articulation disorder, their consonants are often unstable and that makes their voices unclear. Hence, by replacing the articulation-disordered basis of consonants only, a voice with an articulation disorder is converted into a clear voice that preserves the individuality of speaker's voice.

The rest of this paper is organized as follows: In Section 2, NMF-based VC is described, the experimental data is evaluated in Section 3, and the final section is devoted to our conclusions.

2. Voice Conversion Based on NMF

2.1. Basic approach of Exemplar-Based Voice Conversion

In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases.

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \quad (1)$$

$\mathbf{x}_l$ is the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the j-th basis and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ are the collection of the bases and the stack of weights. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights. In this paper, each basis denotes the exemplar of the speech or noise signal, and the collection of exemplar $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called 'dictionary' and 'activity', respectively.

Fig. 1 shows the basic approach of our exemplar-based VC using NMF. D, d, L, and J represent the number of dimensions of source features, dimensions of target features, frames of the dictionary, and basis of the dictionary, respectively. Our VC method needs two dictionaries that are phonemically parallel. One dictionary is a source dictionary, which is constructed from source features. Source features are constructed from an articulation-disordered spectrum and its segment features. The other dictionary is a target dictionary, which is constructed from target features. Target features are mainly constructed from a well-ordered spectrum. These two dictionaries consist of the same words and are aligned with dynamic time warping (DTW). Hence, these dictionaries have the same number of bases.

Input source features X^s, which consist of an articulation-disordered spectrum and its segment features, are decomposed into a linear combination of bases from the source dictionary A^s by NMF. The weights of the bases are estimated as an activity H^s. Therefore, the activity includes the weight information of input features for each basis.

Then, the activity is multiplied by a target dictionary in order to obtain converted spectral features $\hat{X}^t$ which are represented by a linear combination of bases from the target dictionary. Because the source and target dictionary are parallel phonemically, the bases used in the converted features is phonemically the same as that of the source features.

Fig. 2 shows an example of the activity matrices estimated from a word "ikioi" ("vigor" in English). One is uttered by a person with an articulation disorder, and the other is uttered by a physically unimpaired person. To show an intelligible example, each dictionary was structured from just the one word "ikioi" and aligned with DTW. As shown in Fig. 2, these activities have high energies at similar elements. For this reason, when there are parallel dictionaries, the activity of the source features estimated with the source dictionary may be able to be substituted with that of the target features. Therefore, the target speech can be constructed using the target dictionary and the activity of the source signal as shown in Fig. 1.

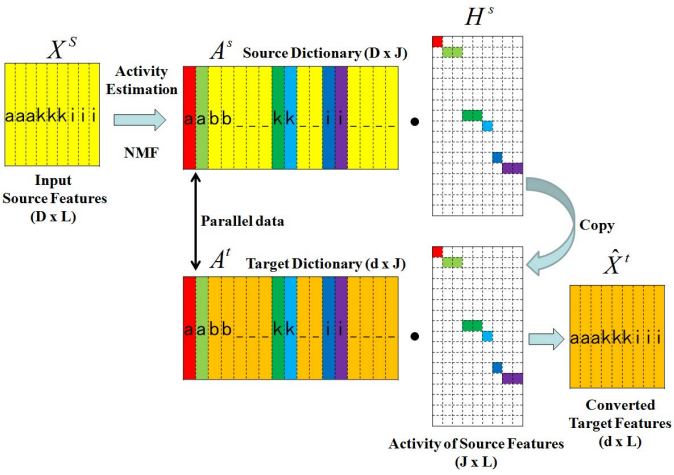

Figure 1: Basic approach of NMF-based voice conversion

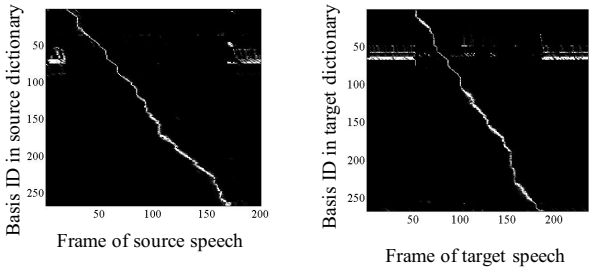

Figure 2: Activity matrices for the articulation-disordered utterance (left) and well-ordered utterance (right)

2.2. Constructing the Individuality-Preserving Dictionary

In the preceding section, both dictionaries (source and target) consisted of the same spectral envelope features extracted by STRAIGHT analysis [11] for simplicity in explaining the proposed method. Indeed, the use of these features worked without any problems in a preliminary experiment using speech data. However, when it came to constructing a noise dictionary, STRAIGHT analysis could not express the noise spectrum well since STRAIGHT itself is an analysis and synthesis method for speech data. In order to express the noisy source speech with a sparse representation of clean source and noise dictionaries, a simple magnitude spectrum calculated by short-time Fourier transform (STFT) is used to construct the source and noise dictionaries.

In order to make a parallel dictionary, some pairs of parallel utterances are needed, where each pair consists of the same text. One is spoken by a person with an articulation disorder (source speaker), and the other is spoken by a physically unimpaired person (target speaker). The left side of Fig. 3 shows the process for constructing a parallel dictionary. The magnitude spectrum is calculated from the clean source utterance to construct source dictionary. For the target dictionary, STRAIGHT spectrum is extracted from clean parallel utterances. The extracted magnitude spectrum and spectrum envelopes are phonemically aligned with DTW. In order to estimate the activities of the source features precisely, segment features, which consist of some consecutive frames, are constructed. Target features are constructed from consonant frames of the target's aligned spectrum and vowel frames of the source's aligned spectrum. Source and target dictionaries are constructed by lining up each of the features extracted from parallel utterances.

The right side of Fig. 3 shows how to preserve a source speaker's voice individuality in our VC method. K represents the number of noise dictionary frames. The vowels of a person's voice strongly imply a speaker's individuality. On the other hand, the consonants of people with articulation disorders are often unstable. By combining a source speaker's vowels and target speaker's consonants in the target dictionary, the individuality of the source speaker's voice can be preserved.

2.3. Estimation of Activity from Noisy Source Signal

From the before- and after-utterance sections in the observed (noisy) signal, the noise dictionary is extracted for each utterance. In the exemplar-based approach, the spectrum of the noisy source signal at frame l is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities.

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\ &\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{a}_k^n h_{k,l}^n \\ &= [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\ &= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \tag{2}$$

$\mathbf{x}_l^s$ and $\mathbf{x}_l^n$ are the magnitude spectra of the source signal and the noise. A^s, A^n, h_l^s, h_l^n are the source dictionary, noise dictionary, and their activities at frame l. Given the spectrogram, (2) can be written as follows:

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{A}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{H}^s \\ \mathbf{H}^n \end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \tag{3}$$

In order to consider only the shape of the spectrum, X, A^s and A^n are first normalized for each frame or exemplar so that the sum of the magnitudes over frequency bins equals unity.

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \mathbf{X} &\leftarrow \mathbf{X}./\mathbf{M} \\ \mathbf{A} &\leftarrow \mathbf{A}./(\mathbf{1}^{(D\times D)}\mathbf{A}) \end{aligned} \tag{4}$$

$\mathbf{1}$ is an all-one matrix and $./$ denotes element-wise division, respectively. The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function [9]:

$$d(\mathbf{X}, \mathbf{AH}) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{5}$$

The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}$ and $\mathbf{AH}$. The second term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{H}$ to be sparse. $.*$ denotes element-wise multiplication. The weights of the sparsity constraints can be defined for each exemplar by defining $\lambda^T = [\lambda_1 \dots \lambda_J \dots \lambda_{J+K}]$. In this paper, the weights for source exemplars $[\lambda_1 \dots \lambda_J]$ were set to 1, and those for noise exemplars $[\lambda_{J+1} \dots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (5) is estimated iteratively applying the following update rule:

$$\begin{aligned}\mathbf{H}_{n+1} &= \mathbf{H}_n .* (\mathbf{A}^T(\mathbf{X}./(\mathbf{AH}))) \\ &\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda \mathbf{1}^{(1\times L)}). \end{aligned} \tag{6}$$

2.4. Target Speech Construction

From the estimated joint matrix $\mathbf{H}$, the activity of source signal $\mathbf{H}^s$ is extracted, and by using the activity and the target dictionary, the converted spectral features are constructed. Then, the target dictionary is also normalized for each frame in the same way the source dictionary was.

$$\mathbf{A}^t \leftarrow \mathbf{A}^t./(\mathbf{1}^{(d\times d)}\mathbf{A}^t) \tag{7}$$

Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (4) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) .* \mathbf{M} \tag{8}$$

The input source feature is the magnitude spectrum calculated by STFT, but the converted spectral feature is expressed as a STRAIGHT spectrum. Hence, the target speech is synthesized using a STRAIGHT synthesizer. The other features extracted by STRAIGHT analysis, such as F0 and the aperiodic components, are used to synthesize the converted signal without any conversion.

3. Experimental Results

3.1. Experimental Conditions

The proposed method was evaluated on word-based VC for one person with an articulation disorder. We recorded 432 utterances (216 words, repeating each two times) included in the ATR Japanese speech database. The speech signals were sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. A physically unimpaired Japanese male in the ATR Japanese speech database was chosen as a target speaker. Two hundred sixteen utterances were used for training, and the other 216 utterances were used for the test. The number of dimensions of source and target features are, 2565 and 513. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database) to the clean speech sentences. The mean SNR was about 20 dB. The noise dictionary is extracted from the before- and after-utterance section in the evaluation sentence.

We compared our NMF-based VC to conventional GMM-based VC. In GMM-based VC, the 1st through 24th cepstrum coefficients extracted by STRAIGHT were used as source and target features.

3.2. Subjective Evaluation

We conducted subjective evaluation on 3 topics. A total of 5 Japanese speakers performed the test using headphones. For the "listening intelligibility" evaluation, we performed a MOS

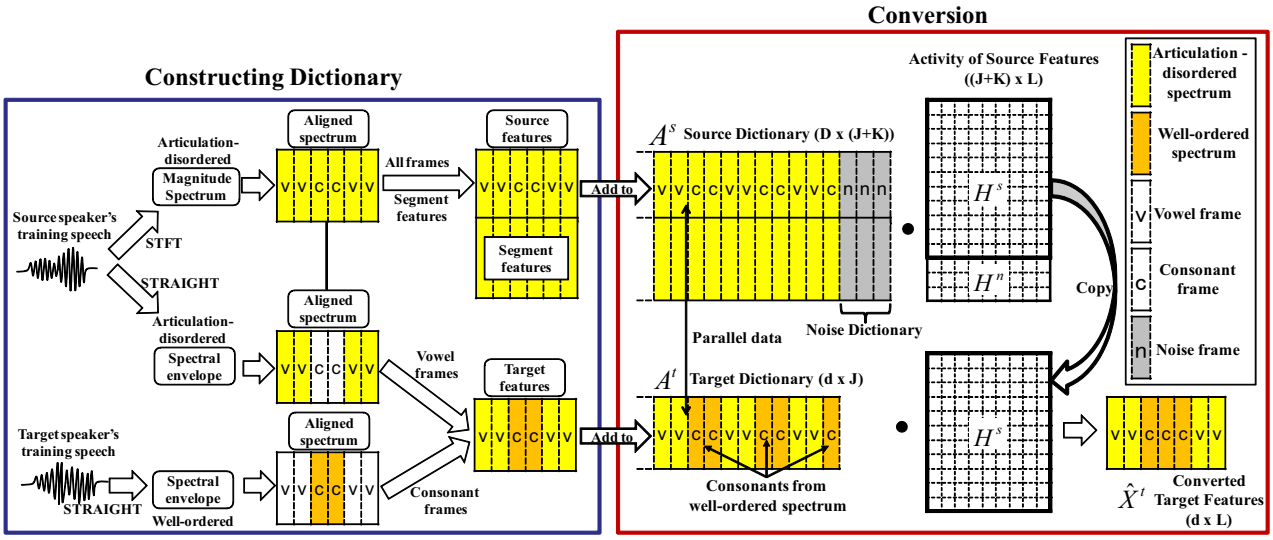

Figure 3: Individuality-preserving voice conversion

(Mean Opinion Score) test. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Twenty words, which are difficult for a person with articulation disorder to utter, were evaluated. The subjects were asked about the listening intelligibility in the articulation-disordered voice, the NMF-based converted voice, and the GMM-based converted voice. Each voice uttered by a physically unimpaired person was presented as a reference of 5 points on the MOS test.

Fifty words were converted using NMF-based VC and GMM-based VC for the following evaluations. On the "similarity" evaluation, the XAB test was carried out. In the XAB test, each subject listened to the articulation disordered speech. Then the subject listened to the speech converted by the two methods and selected which sample sounded more similar to the articulation disordered speech. On the "naturalness" evaluation, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods and selected which sample sounded more natural.

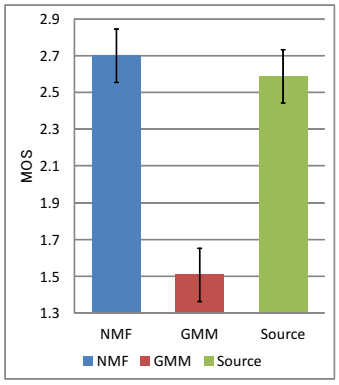

Figure 4: Results of MOS test on listening intelligibility

Fig. 4 shows the results of the MOS test for listening intelligibility. The error bars show a 95% confidence score. As shown in Fig. 4, NMF-based VC can improve listening intelligibility. On the other hand, GMM-based conversion deteriorates

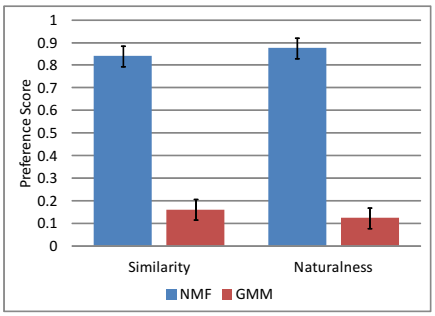

Figure 5: Preference scores for the similarity to the source speaker and naturalness

the listening intelligibility. This might be because background noise caused unexpected mapping in the GMM-based VC and degraded the conversion performance.

Fig. 5 shows the preference score on the similarity to the source speaker and naturalness of the converted voice. NMF-based VC got a higher score than GMM-based conversion on similarity because NMF-based conversion used a combined dictionary. NMF-based VC also got a higher score than GMM-based conversion on naturalness although NMF-based conversion mixed the source speaker's vowels and target speaker's consonants.

4. Conclusions

We proposed a noise robust spectral conversion method based on NMF for a voice with an articulation disorder. Experimental results demonstrated that our VC method can improve the listening intelligibility of words uttered by a person with an articulation disorder in noisy environments. Moreover, compared to conventional GMM-based VC, NMF-based VC can preserve the individuality of the source speaker's voice and the naturalness of the voice. In this study, there was only one subject person, so in future experiments, we will increase the number of subjects and further examine the effectiveness of our method.

5. References

[1] S. T. Canale and W. C. Campbell, "Campbell's operative orthopaedics," Mosby-Year Book, Tech. Rep., 2002.

[2] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," *Proc. Interspeech*, 2012.

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

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

[5] Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *IEEE Trans. Seech and Audio Proc., Vol. 7, pp. 2401-2404*, 1999.

[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing, Vol. 2 No. 5*, 2012.

[7] 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.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *in Proc. Neural Information Processing System, pp. 556-562*, 2001.

[9] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," *ICASSP, pp. 4546-4549*, 2010.

[10] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," *INTERSPEECH*, 2006.

[11] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication, Vol. 27, No. 3-4*, 1999.