Exemplar-based unit selection for voice conversion utilizing temporal information

Zhizheng Wu[1], *Tuomas Virtanen*[2], *Tomi Kinnunen*[3], *Eng Siong Chng*[1], *Haizhou Li*[1,4]

[1]Nanyang Technological University, Singapore
[2]Tampere University of Technology, Tampere, Finland
[3]University of Eastern Finland, Joensuu, Finland
[4]Institute for Infocomm Research, Singapore

wuzz@ntu.edu.sg

Abstract

Although temporal information of speech has been shown to play an important role in perception, most of the voice conversion approaches assume the speech frames are independent of each other, thereby ignoring the temporal information. In this study, we improve conventional unit selection approach by using exemplars which span multiple frames as base units, and also take temporal information constraint into voice conversion by using overlapping frames to generate speech parameters. This approach thus provides more stable concatenation cost and avoids discontinuity problem in conventional unit selection approach. The proposed method also keeps away from the over-smoothing problem in the mainstream joint density Gaussian mixture model (JD-GMM) based conversion method by directly using target speaker's training data for synthesizing the converted speech. Both objective and subjective evaluations indicate that our proposed method outperforms JD-GMM and conventional unit selection methods.

Index Terms: Voice conversion, unit selection, multi-frame exemplar, temporal information

1. Introduction

The task of *voice conversion* is to modify one speaker's voice (source) to sound like another (target). It has many applications in unit selection based speech synthesis, such as personalization of a text-to-speech (TTS) system without the need to retrain a full TTS system for each target speaker [1]. To be useful in such applications, natural sounding and high quality speech generated from the voice conversion system is expected.

A number of methods have been proposed in order to generate natural sounding converted speech. One of the successful methods is to estimate a parametric conversion function from a parallel training corpus, and then to apply this conversion function to convert the unseen test utterances. For instance, methods such as *joint density Gaussian mixture model* (JD-GMM) [2, 3], *partial least squares regression* [4], *mixture of factor analyzers* [5] and *local linear transformation* [6] have been studied making use of local linear transformation functions. Nonlinear mapping approaches such as neural network [7, 8], dynamic kernel partial least squares regression [9] and conditional restricted Boltzmann machine [10] have also been proposed, assuming that the vocal tract shape differences between two speakers constitute a non-linear relationship. All of the above methods can generate converted speech with acceptable quality. However, *over-smoothing* and *over-fitting* problems in these statistical methods have been reported in [11, 9, 6, 5], due to statistical average and large number of parameters, respectively, and these problems affect the quality of synthesized speech considerably.

Without using transformation functions, it is also possible to directly utilize the original target speech parameters to generate converted speech. *Unit selection* [12], a method of automatically selecting and concatenating target speech segments, is a representative example of such non-parametric methods. In [13], unit selection method, which uses source speech as reference speech for selecting the target units, is proposed for text-independent voice conversion. In [14], the authors improved the original unit selection approach [13] by using JD-GMM based converted speech as reference speech. To avoid discontinuities at the concatenated boundaries, the unit selection methods [13, 14] consider both the *target cost* and *concatenation cost*. Unfortunately, they only use one frame to calculate the concatenation cost, which has not considered a smooth frame-to-frame transition in the target space. In addition, temporal information is also ignored in the generated speech parameter sequence, which will result in the discontinuity at the concatenation points and affect the perceptional quality of the synthesized speech.

A major concern in most of the conventional voice conversion methods is that they assume the short-term frames are independent observations of each other. Inspired by the findings in exemplar-based speech recognition [15] which considers the dependency of multiple frames, we propose an exemplar-based unit selection method to avoid frame-by-frame independence assumption. We use exemplars which span over a fixed number of frames as basic units to calculate the concatenation cost and to generate converted speech parameters to avoid discontinuity at the concatenation boundaries. Compared with the previous unit selection approaches [13, 14], our method has three novel contributions:

a) we use a multi-frame exemplar instead of a single frame as basic unit;

b) we adopt the exemplars to calculate the concatenation cost to ensure that the consecutive frames in the target space have zero concatenation cost;

c) we utilize the temporal information constraint to generate the converted speech parameters by using a temporal window to deal with the overlapping frames between consecutive exemplars.

In contrast to the statistical approaches, our method directly makes use of the target speaker's training data to generate the converted speech, which will avoid both over-smoothing and over-fitting problems.

We summarize the process as follows. We first find source-target exemplar pairs on parallel training data; we then select several target candidate exemplars for each source exemplar in a test sentence and calculate the target cost and the concatenation cost; after that Viterbi algorithm is adopted to find the optimal target exemplar sequence which minimizes the overall target and concatenation costs; finally, the converted speech parameters are generated from the overlapping exemplars by considering temporal information constraint.

2. Exemplar-based Unit Selection

An *exemplar* is a time-frequency speech segment which spans over multiple consecutive frames. Exemplar-based methods have been popular in modern speech recognition [15], as they allow modelling of the temporal information. Different from the template-based speech recognition [16] and unit selection for concatenative speech synthesis [12] which employ transcription label to obtain the template or unit, we use exemplars with fixed number of frames similar as [17], because the transcription information is not available in this study.

2.1. Source-target exemplars pairing

Given a parallel data, source frame sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_x}, ..., \mathbf{x}_{N_x}]$ and target frame sequence $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{n_y}, ..., \mathbf{y}_{N_y}]$, dynamic time warping (DTW) is performed to obtain aligned frames. The alignment produces the joint vector sequence $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N]$, where $\mathbf{z}_n = [\mathbf{x}_{n_x}^\top, \mathbf{y}_{n_y}^\top]^\top$, $\mathbf{x}_{n_x} \in \mathcal{R}^D, \mathbf{y}_{n_y} \in \mathcal{R}^D$ and $\mathbf{z}_n \in \mathcal{R}^{2D}$. Hence, the exemplar pair at time n is

$$\mathbf{X}^{(n)} = [\mathbf{x}_{n_x-p}, \mathbf{x}_{n_x-p+1}, ..., \mathbf{x}_{n_x}, ..., \mathbf{x}_{n_x+p-1}, \mathbf{x}_{n_x+p}] \in \mathcal{R}^{q\times D}$$

for source and

$$\mathbf{Y}^{(n)} = [\mathbf{y}_{n_y-p}, \mathbf{y}_{n_y-p+1}, ..., \mathbf{y}_{n_y}, ..., \mathbf{y}_{n_y+p-1}, \mathbf{y}_{n_y+p}] \in \mathcal{R}^{q\times D}$$

for target, where $q = 2p + 1$ is the window size of an exemplar. We note that two consecutive exemplars $\mathbf{X}^{(n)}$ and $\mathbf{X}^{(n+1)}$ have $(q - 1)$ overlapping frames. We note that there are no repeated frames within an exemplar.

2.2. Pre-selection of candidate exemplars

Note that we obtain exemplar pairs from parallel training data. At run-time testing, for each source exemplar $\mathbf{X}^t = [\mathbf{x}_{t-p}, \mathbf{x}_{t-p+1}, ..., \mathbf{x}_t, ..., \mathbf{x}_{t+p-1}, \mathbf{x}_{t+p}]$ in the testing sentence, we pre-select several target exemplars as candidates.

We first find the K nearest neighbors $\mathbf{X}_1^{'(t)}, ..., \mathbf{X}_K^{'(t)}$ in the source training data for each $\mathbf{X}^{(t)}$. The paired target exemplars $\mathbf{Y}_1^{'(t)}, ..., \mathbf{Y}_K^{'(t)}$ corresponding to $\mathbf{X}_1^{'(t)}, ..., \mathbf{X}_K^{'(t)}$ are then selected based on the source-target exemplars pairing in the previous step. Thus, the **target cost** for each candidate is calculated as follows:

$$\mathcal{C}_{\text{target}}(\mathbf{X}^{(t)}, \mathbf{Y}_k^{'(t)}) = \sum_{i=1}^{q}\sum_{d=1}^{D}(\mathbf{X}^{(t)}(i,d) - \mathbf{Y}_k^{'(t)}(i,d))^2, \tag{1}$$

where $\mathbf{X}^{(t)}(i,d)$ and $\mathbf{Y}_k^{'(t)}(i,d)$ are the d-th dimension elements of the i-th frame vector of exemplars $\mathbf{X}^{(t)}$ and $\mathbf{Y}_k^{'(t)}$ at time t, respectively.

After the shortlisted candidate exemplars are chosen, we calculate the target-to-target **concatenation cost** as follows:

$$\mathcal{C}_{\text{concatenation}}(\mathbf{Y}_k^{'(t)}, \mathbf{Y}_j^{'(t+1)}) = \sum_{l=1}^{q-1}\sum_{d=1}^{D}(\mathbf{Y}_k^{'(t)}(l+1,d) - \mathbf{Y}_j^{'(t+1)}(l,d))^2; j = 1,..,K \tag{2}$$

where $\mathbf{Y}_k^{'(t)}(l+1,d)$ is the d-th dimension element of $(l+1)$-th frame vector of the k-th candidate at time t. We note that if two exemplars are exactly the neighbours in the training set, the concatenation cost will be 0, because the $(q-1)$ frames used for calculation are exactly the same.

If the window size is one $(q = 1)$, $\mathbf{Y}_k^{'(t)}$ becomes a D-dimensional vector. In this special case, the concatenation cost is:

$$\mathcal{C}_{\text{concatenation}}(\mathbf{Y}_k^{'(t)}, \mathbf{Y}_j^{'(t+1)}) = \sum_{d=1}^{D}(\mathbf{Y}_k^{'(t)}(d) - \mathbf{Y}_j^{'(t+1)}(d))^2; j = 1,...,K. \tag{3}$$

This is the same as the calculation of conventional concatenation cost, which can not guarantee the cost to be 0 when the two frames are exactly neighbours, as two consecutive frames may not be exactly the same.

2.3. Searching for the optimal exemplar sequence

Given a source exemplar sequence $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(t)}, ..., \mathbf{X}^{(T)}$ from a testing sentences, K target exemplars for each source exemplar are pre-selected, and the target cost and the concatenation cost are all calculated as introduced in the previous step. Then, the optimal target exemplar sequence can be found by minimizing the following cost function:

$$\tilde{\mathbf{Y}}^{(1)}, ..., \tilde{\mathbf{Y}}^{(T)} = \arg\min_{k=1,...,K}\sum_{t=1}^{T}\{\mathcal{C}_{\text{target}}(\mathbf{X}^{(t)}, \mathbf{Y}_k^{'(t)}) + \mathcal{C}_{\text{concatenation}}(\mathbf{Y}_k^{'(t)}, \mathbf{Y}_j^{'(t+1)})\}; j = 1, ..., K. \tag{4}$$

In practice, this is achieved by using Viterbi search, as illustrated in Fig. 1.

2.4. Speech parameter generation

Although the exemplar sequence is obtained using Viterbi search, we can not directly pass the exemplar sequence to the synthesis filter to reconstruct the speech signal, because there are overlapping frames between consecutive exemplars. The overlapping frames contain temporal information that is beneficial for a smooth signal re-construction. To take advantage of such temporal information, we introduce a weight for each frame in an exemplar, which forms a temporal window.

$$\mathbf{a} = [p, p-1, ..., 0, ..., p-1, p], \tag{5}$$

$$\mathbf{w} = \exp(-\lambda|\mathbf{a}|), \tag{6}$$

where λ is a scalar value to control the shape of the temporal window. We normalize the weight vector $\mathbf{w}$ to make sure the

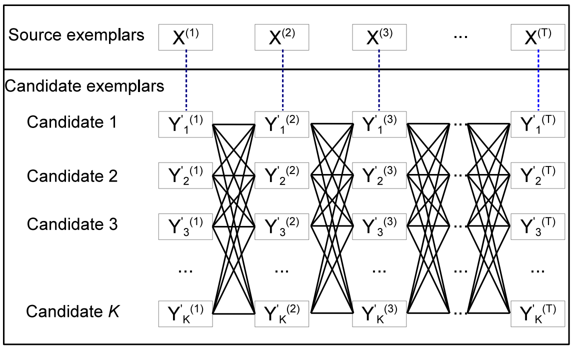


Figure 1: Illustration of searching for the optimal exemplar sequence. In the figure, dashed line (connecting $\mathbf{X}^{(t)}$ and $\mathbf{Y}_k^{'(t)}$) represents *target cost* and solid line (connecting $\mathbf{Y}_k^{'(t)}$ and $\mathbf{Y}_j^{'(t+1)}$) represents *concatenation cost*.

elements sum to 1. The converted speech parameters are generated as follows:

$$\mathbf{y}^{\tilde{(t)}} = \sum_{i=1}^{q} \tilde{\mathbf{Y}}^{(t-p+i-1)}(q-i+1) \times \mathbf{w}(q-i+1) \qquad (7)$$

where $\tilde{\mathbf{Y}}^{(t-p+i-1)}(q-i+1)$ is the $(q-i+1)$-th column vector of $\tilde{\mathbf{Y}}^{(t-p+i-1)}$, and $\mathbf{w}(q-i+1)$ is the $(q-i+1)$-th element of $\mathbf{w}$.

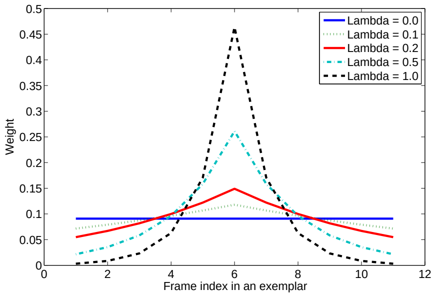


Figure 2: The temporal window with different λ

Fig. 2 shows the shape of the temporal window for different values of λ. As we increase λ, the contribution of the center frame to the converted speech parameter increases as well. While choosing large enough λ is similar as choosing only the center frame as the converted speech parameter. Conversely, decreasing λ means considering more temporal information, and $\lambda = 0.0$ corresponds to merely averaging all the overlapped frames.

If we set the center element of $\mathbf{w}$ to one and the rest elements to zero, in other words, only the center frame in an exemplar is used as the converted speech parameter, this will reduce the method to a scheme which does not take into consideration of temporal information constraint in the synthesis and only in finding the optimal sequence of exemplars.

3. Experiments

The CMU ARCTIC corpus is adopted for the experiments. Two male (BDL, RMS) and two female (SLT, CLB) speakers are selected. 200 utterances of each speaker are used as training data, and 20 utterances of each speaker are used as testing data. We conduct both inter-gender and intra-gender conversions: BDL to RMS (M2M), BDL to SLT (M2F), SLT to CLB (F2F) and SLT to RMS (F2M).

The speech signal, sampled at 16 kHz, is analyzed using STRAIGHT [18] with 5ms shift. 24-order mel-cepstral coefficients (MCC), excluding the 0th energy coefficient, are extracted. The MCCs are converted by using each of the conversion method detailed in the following paragraph, while log-scale F0 is converted by equalizing the means and variances of the source and the target speakers.

In this work, we compare the following four approaches:

a) *Joint density Gaussian mixture model (JD-GMM)*: This is the mainstream voice conversion method [2, 3]. We adopt 64 full covariance Gaussian components to model the joint distribution of source and target speech. This is our first baseline method.

b) *Unit selection (US)* [13, 14]: This is the conventional unit selection approach, using only one frame to calculate both the target and the concatenation costs. This is our second baseline method.

c) *Partial exemplar-based unit selection (PEUS)*: The method follows the steps as described in section 2.1, 2.2 and 2.3. While in the generation step, only the center frame in an exemplar is chosen to generate the converted speech parameters. It is an intermediate method towards our proposed method.

d) *Exemplar-based unit selection (EUS)*: This method is the proposed method as detailed in previous section.

3.1. Objective evaluation

To evaluate the performance objectively, we adopt mel-cepstral distortion [2, 3] as an objective evaluation measure:

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\Sigma_{i=0}^{24}(mc_i^t - mc_i^c)^2}, \qquad (8)$$

where mc^t and mc^c are the target and converted MCCs, respectively. The lower of the MCD value, the smaller distortion.

We first study the effects of the window size of an exemplar (q) and the number of shortlisted candidates (K). Here we only use the center frame in the exemplar without any overlapping constraints to generate converted speech.

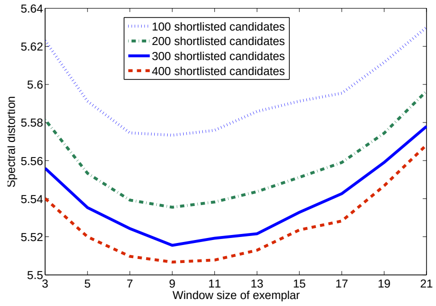


Figure 3: Spectral distortion as a function of window size of an exemplar and number of shortlist candidates in terms of spectral distortion (dB).

Fig. 3 indicates that window size of $q = 9$ can give lowest distortion consistently. Distortion decreases with increased shortlist size as expected, but comes with an added computational overhead. Since there is not much change beyond 200

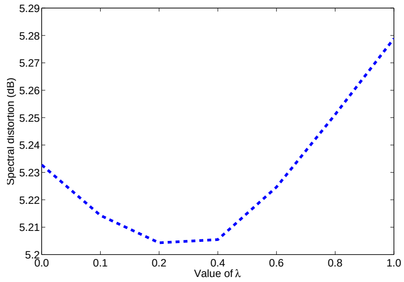


Figure 4: Distortion as a function of λ.

candidates, we fix $K = 200$ and $q = 9$ for the rest of the experiments.

We now turn our attention to the temporal window. As shown in Eq. (6), we use λ to control the shape of the temporal window for an exemplar. The distortion with different values of λ are shown in Fig. 4. When $\lambda \geq 0.4$, the distortion increases. When $\lambda < 0.2$, the distortion also increases. Therefore, we empirically fix $\lambda = 0.2$ for the rest of the experiments.

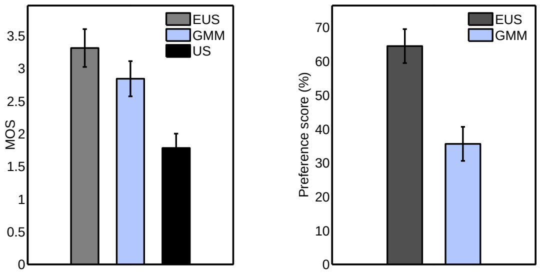
Table 1: Spectral distortion comparison of the baselines and the proposed exemplar-based unit selection (EUS) method

	M2M	M2F	F2F	F2M	Average
JD-GMM	5.28	5.60	4.82	5.34	5.26
US	6.25	6.69	5.62	6.39	6.24
PEUS	5.57	5.90	5.04	5.52	5.51
EUS	**5.26**	**5.57**	**4.77**	**5.21**	**5.20**

Comparison of the proposed method with the two baseline methods and PEUS method is given in Table 1. Comparing with US and PEUS, PEUS gives lower spectral distortion and we can see the advantage of using multiple frames exemplar to do preselection and to calculate the target cost and concatenation cost for searching the optimal frame sequence. The benefit of using a temporal window to include temporal information constraint in the converted speech parameter generation can be seen by comparing the results of PEUS and EUS. The difference of PEUS and EUS methods is that EUS employs a temporal window to deal with overlapping frames between consecutive exemplars while PEUS does not. JD-GMM method also gives higher spectral distortion than the proposed EUS method for both male and female source speakers. In general, the proposed method (EUS) has lower spectral distortion than both JD-GMM and US methods. We note again that the lower spectral distortion, the better performance.

3.2. Subjective evaluation

To assess the overall quality of converted speech, we conducted subjective evaluation using mean opinion score (MOS). We compare the proposed EUS method with the two baseline methods: JD-GMM and US. As PEUS is an intermediate method towards EUS, it is excluded in the subjective evaluation. We randomly select 5 sentences from JD-GMM conversion, US conversion and EUS conversion speech of four conversion directions (M2M, M2F, F2F and F2M). As a result, there are 20 sentences for each method and 60 sentences in the whole test. These speech samples were presented to 9 subjects. The subjects were asked to listen to each speech sample and then rate the speech quality based on a five point scale: 5 for *perfect*, 4 for *good*, 3 for *fair*, 2 for *poor*, and 1 for *bad*. The MOS is obtained by average all the scores rated by all the subjects. The MOS results are presented in Fig. 5(a). We can see that our proposed method outperforms both JD-GMM and conventional unit selection method in terms of perceptual quality. The proposed temporal window in EUS method is able to smooth the converted trajectory, while the US method without such temporal window can not.



(a) Overall quality results

(b) Similarity results

Figure 5: Subjective evaluation results with 95% confidence interval

An AB preference test was also conducted to access the similarity of the converted speech. As the bad quality of US converted speech affect the similarity test, we only compare the proposed EUS method with baseline JD-GMM method. 9 subjects were asked to listen to a reference target speech and one pair converted speech (A and B), and decide to choose which speech sample, A or B, is more similar to the reference target speech. The preference results are shown in Fig. 5(b). It clearly shows that the proposed EUS method can generate speech to sound more similar to the target speaker than the conventional JD-GMM method. We note that EUS method directly select target frames to compose the converted speech. Thus, it is easy to understand that it generates speech more similar to target speaker than JD-GMM based conversion does, the latter employs a transformation function to transform the source speech to the target space.

4. Conclusions

In this paper, to avoid the frame-by-frame independence assumption in most the voice conversion methods, we proposed exemplar-based unit selection method to model the temporal dependency and take into consideration of temporal information constraint in both the process of finding the optimal exemplar sequence and generation of converted speech parameters. By using multi-frame exemplars, the proposed method avoids the discontinuity at the concatenation point in conventional unit selection approaches. In addition, our method also avoids the over-smoothing problem in the popular JD-GMM approach because the target speaker's training data is directly used to generate the converted speech. Generally, the proposed method has lower spectral distortion, and also generates perceptually better speech than baseline methods.

5. Acknowledgement

This work was supported by Academy of Finland (projects no. 132129, 253120, 258708). The authors would like to thank all the listeners who take part in the subjective evaluation test.

6. References

[1] Y. Stylianou, "Voice transformation: a survey," in *ICASSP 2009*.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

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

[4] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[5] Z. Wu, T. Kinnunen, E. Chng, and H. Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *Signal Processing Letters, IEEE*, vol. 19, no. 12, pp. 914–917, 2012.

[6] V. Popa, H. Silen, J. Nurminen, and M. Gabbouj, "Local linear transformation for voice conversion," in *ICASSP 2012*.

[7] M. Narendranath, H.A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[8] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[9] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[10] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) 2013*.

[11] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Eurospeech-2003*.

[12] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP 1996*.

[13] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *ICASSP 2006*.

[14] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *ICASSP 2007*.

[15] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J.F. Gemmeke, J.R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98 –113, nov. 2012.

[16] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1377–1390, 2007.

[17] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring 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*, vol. 27, no. 3, pp. 187–207, 1999.