

# Voice Conversion Methods for Vocal Tract and Pitch Contour Modification

Oytun Turk

R&D Dept.,  
Sestek Inc., Istanbul, Turkey  
oytun@sestek.com.tr

Levent M. Arslan

Electrical and Electronics Eng. Dept.,  
Bogazici University, Istanbul, Turkey  
arslanle@boun.edu.tr

## Abstract

This study<sup>1</sup> proposes two new methods for detailed modeling and transformation of the vocal tract spectrum and the pitch contour. The first method (selective pre-emphasis) relies on band-pass filtering to perform vocal tract transformation. The second method (segmental pitch contour model) focuses on a more detailed modeling of pitch contours. Both methods are utilized in the design of a voice conversion algorithm based on codebook mapping. We compare them with existing vocal tract and pitch contour transformation methods and acoustic feature transplantations in subjective tests. The performance of the selective pre-emphasis based method is similar to the methods used in our previous work at higher sampling rates with a lower prediction order. The results also indicate that the segmental pitch contour model improves voice conversion performance.

## 1. Introduction

Detailed modeling and transformation of the vocal tract spectrum and the pitch contour are two key issues for the design of voice conversion algorithms. As we have shown in our previous work, vocal tract and pitch characteristics have a dominant role in perception of speaker identity [1]. Several methods are used for modeling and transformation of the vocal tract spectrum. Examples include formant frequencies [2], and the sinusoidal model parameters [3]. Line spectral frequencies (LSFs) attract special attention because of their nice interpolation properties as described in [4], and [5]. Several methods for pitch contour modeling and transformation are described in [6].

This study proposes two new methods for detailed modeling and transformation of the vocal tract spectrum and the pitch contour. In Section 2, we describe the selective pre-emphasis method to model and transform the vocal tract spectrum. We employ a sub-band based framework taking the perceptual characteristics of the human auditory system into account. The selective pre-emphasis method provides the means for detailed spectral envelope estimation and for modification of spectral resolution in different sub-bands. In Section 3, we propose a segmental pitch contour model. Both methods are incorporated into the voice conversion algorithm of STASC [4]. Section 4 describes a subjective test for the evaluation of new methods. Finally, in Section 5, the results and future work are discussed.

## 2. Selective Pre-emphasis System

It is common practice to apply pre-emphasis prior to LPC analysis to enhance the numerical properties of the procedure. In this section, we combine the motivation behind pre-

emphasis with perceptual sub-band processing to estimate the vocal tract spectrum in detail. We refer to this new method as selective pre-emphasis. The selective pre-emphasis system was developed to overcome the problems of the Discrete Wavelet Transform (DWT) based system that performs transformation of the vocal tract spectrum in different sub-bands as described in [7]. Although it provides efficient solutions at higher sampling rates, it has certain disadvantages. When the aim is to perform modification in all sub-bands, aliasing distortion causes a reduction in the output quality. This is due to the fact that modified version of one sub-band may overlap with another sub-band in the reconstruction stage. For this reason, we have investigated a new method that provides the means to model and transform different frequency regions in different amounts of spectral detail with less distortion and more flexibility. We use the fact that LPC analysis models spectral peaks better than spectral nulls. So, it is possible to emphasize specific regions of the spectrum by band-pass filtering to capture spectral details.

### 2.1. Analysis and Synthesis

The basic idea behind selective pre-emphasis is to estimate the vocal tract spectrum as a weighted combination of the spectral envelopes of the sub-band components. First, the speech signal  $s(n)$  is filtered with a bandpass filterbank. Each sub-band component is processed frame-by-frame. LP analysis is performed on each sub-band component to obtain  $a_i(r)$ , the LP coefficients of the  $i^{\text{th}}$  sub-band component, and  $H_i(k)$ , the spectral envelope. Next,  $H(k)$ , the full-band vocal tract spectrum is calculated using Equation 1 as a weighted combination of the sub-band spectral envelopes. We denote the weight of the LP spectrum of a sub-band component at a specific frequency  $k$  by  $c_i(k)$  as given by Equation 2. Note that,  $k_1$  is the lower cut-off frequency of the  $(i+1)^{\text{th}}$  band-pass filter, and  $k_2$  is the higher cut-off frequency of the  $i^{\text{th}}$  band-pass filter. The condition  $k_1 \leq k \leq k_2$  ensures that the band-pass filters are overlapping. The flowchart for selective pre-emphasis based analysis is shown in Fig. 1.

$$H(k) = \sum_i c_i(k) H_i(k) \quad (1)$$

$$c_i(k) = \begin{cases} \frac{k - k_1}{k_1 - k_2} + 1, & \text{if } k_1 \leq k \leq k_2 \\ 0, & \text{else} \end{cases} \quad (2)$$

In the synthesis stage, we use the synthesis LP coefficients and synthesis excitation spectra to obtain the output signal. Note that we use the hat symbol for the synthesis parameters because they can be modified versions of the analysis parameters depending on the application. The synthesis stage

<sup>1</sup> This study was supported by Bogazici University 03A201 Research Fund Project.

is the reverse of the analysis stage as shown in Fig. 1. In Fig. 2, we demonstrate the selective pre-emphasis based spectral estimation method using a filterbank with 4 equally spaced sub-bands in the range 0.0-8.0 KHz. Linear prediction order was 18 for both full-band LP and selective pre-emphasis for a sampling rate of 16 KHz. So, more detailed spectral estimation is possible using the selective pre-emphasis method without the need to increase the prediction order.

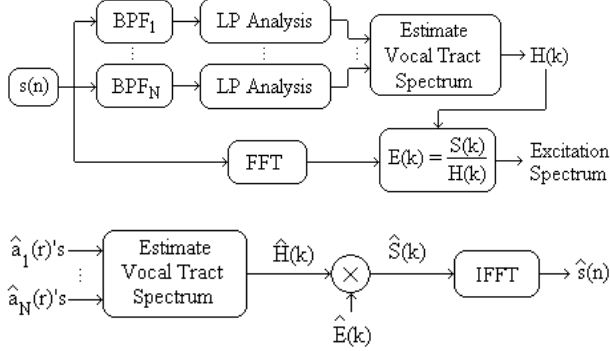


Figure 1: Flowcharts for the analysis algorithm (top) and synthesis algorithm (bottom) for selective pre-emphasis

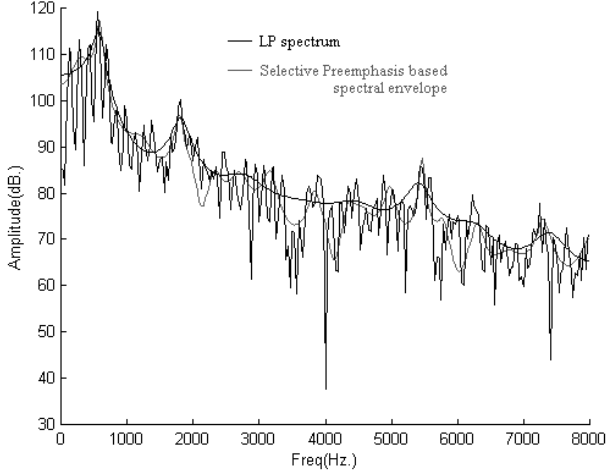


Figure 2: LP vs. selective pre-emphasis based spectral estimation (left), sub-band components and LSFs (right)

Freq. Range (Hz.)	Selective Preemphasis		LP Analysis	
	Mean	Std. Dev.	Mean	Std. Dev.
0-22050	0.43	0.10	0.44	0.11
0-1034	1.00	1.11	1.05	1.18
1034-1895	0.45	0.30	0.54	0.40
1895-2756	0.44	0.24	0.49	0.34
2756-3618	0.40	0.22	0.44	0.26
3618-4823	0.39	0.18	0.44	0.21
4823-6546	0.37	0.12	0.40	0.13
6546-8269	0.38	0.12	0.40	0.13
8269-11714	0.39	0.08	0.39	0.08
11714-15159	0.39	0.08	0.40	0.09
15159-22050	0.42	0.07	0.40	0.07

Table 1: Comparison of LP analysis (P=50) and selective pre-emphasis (P=24) in terms of spectral distances

We have performed an objective test for the comparison of the spectral estimation performance of LP analysis and selective pre-emphasis system for 44.1 KHz signals. In this test, the average spectral distance between the estimated spectrum and the original DFT spectrum is calculated. Two methods are employed for spectral estimation: full-band LP analysis and selective pre-emphasis. Table 1 shows the mean and the standard deviations of the spectral distances. We observe that selective pre-emphasis performs better than LP analysis at a lower prediction order, P.

## 2.2. Training and Transformation

The selective pre-emphasis method is incorporated in the voice conversion algorithm of STASC [4], which consists of two stages: Training and transformation. We have designed a perceptual filterbank for selective pre-emphasis using FIR filters of order 50 as shown in Fig. 3. In the training stage, we use the same utterances of source and target speakers for estimating the acoustical mapping between them. We start with the analysis of the source and target utterances using selective pre-emphasis. An HMM is generated for the lower sub-band components of each source utterance. The target utterance is force-aligned with the source utterance using this HMM. The alignment generated for the lower sub-band components is used for the rest of the sub-bands. Next, we generate codebooks for each sub-band component that contain the line spectral frequencies (LSFs). Fig. 4. shows the flowchart of the training algorithm.

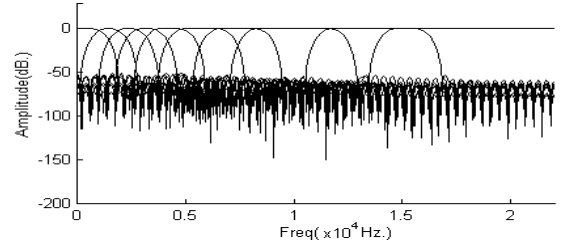


Figure 3: Perceptual filterbank for selective pre-emphasis

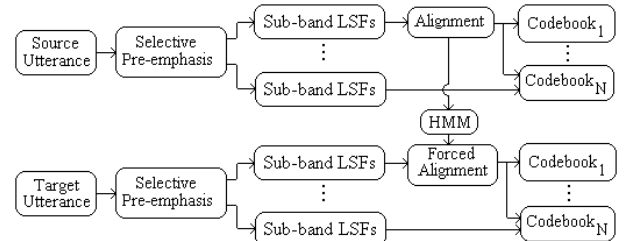


Figure 4: Selective pre-emphasis based training

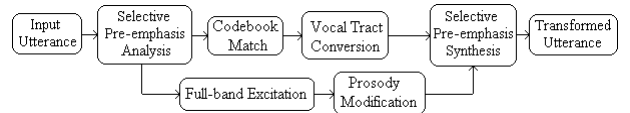


Figure 5: Selective pre-emphasis based transformation

The sub-band codebooks are used for transforming each sub-band of the vocal tract spectrum separately. For this purpose, the input signal is analyzed using selective pre-emphasis. Full-band excitation spectrum is processed

separately for pitch scale modifications. Each sub-band component of the vocal tract spectrum is converted using a weighted average of the corresponding codebook entries as in STASC [4]. Note that the closest codebook entries are estimated using the lower sub-bands and identical entries are used for all sub-bands. Synthesis is performed using the method described in Section 2.1. The flowchart for the transformation algorithm is shown in Fig. 5.

### 3. Segmental Pitch Contour Model

A common approach for modeling pitch is to assume that the pdf of the pitch values is a Gaussian distribution. In this case, it is fairly easy to estimate and transform the pitch values as described in [4] and [6]. However, the local shapes of the pitch contour segments are not modeled and transformed using this approach. To overcome this problem, we have estimated the corresponding pitch contour segments of the source and the target speakers and used this mapping in the transformation stage. We have used identical utterances of the source and the target speakers for training the model. The utterances are aligned; pitch contours are extracted, and smoothed. Target pitch contours are interpolated linearly in the unvoiced parts. Voiced segments of source  $f_0$  contours are extracted. For each voiced source  $f_0$  segment, the corresponding target segment is found using the alignment information. Note that  $s_i$  denotes the  $i^{\text{th}}$  source segment and  $t_i$  is the corresponding target segment. These segments are kept in a pitch contour codebook file. In the transformation stage, the voiced segments,  $f_j$ 's, of the input pitch contour are found. We denote the length of segment  $f_j$  as  $N_j$ . Source and target codebook entries are interpolated to length  $N_j$ . The normalized distance of  $f_j$  to the  $i^{\text{th}}$  source codebook entry is calculated using Equation 3. Next, we estimate a weight for each source codebook segment using Equation 4. We have used  $\alpha=500$  to ensure that only a few close matches from the codebook are included in the generation of the synthetic segment. The synthetic pitch contour segment,  $o_j$ , is estimated using the weights and the target codebook entries using Equation 5. An example for pitch contour transformation using the segmental model is shown in Fig. 6.

$$d_i = \frac{\sum_{n=1}^{N_j} |f_j(n) - s_i(n)|^2}{\sum_{all\ i} \sum_{n=1}^{N_j} |f_j(n) - s_i(n)|^2} \quad (3)$$

$$w_i = \frac{\exp(-\alpha d_i)}{\sum_{all\ i} \exp(-\alpha d_i)} \quad (4) \quad o_j(n) = \sum_{all\ i} w_i t_i(n) \quad (5)$$

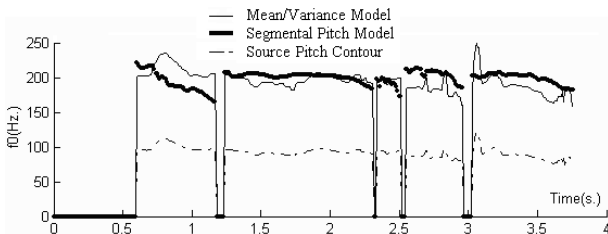


Figure 6: Pitch transformation with the segmental model

## 4. Evaluations

We have designed a subjective test for comparing three vocal tract and two pitch contour transformation methods. The vocal tract conversion methods are the full-band system in STASC with pre-emphasis [4], the DWT based system [7], and the selective pre-emphasis system. For pitch conversion, we have employed the mean/variance model [4], [6] and the segmental pitch contour model. We have used a Turkish database of four male and four female speakers (30 sentences, 50 words, recorded at 44.1 KHz). First, the full-band, DWT and selective pre-emphasis systems are trained separately for each source/target speaker pair. The segmental pitch contour model was trained while performing full-band training. 15 test utterances (5 sentences, 10 words) are transformed using all methods in Table 2. We have also included vocal tract and pitch transplantation outputs in the test for comparison. Note that the transplantations correspond to the ideal case for the transformation of a feature as the exact mapping between the source and target features are known.

VT Conversion	P Conversion	Symbol	Output Type
-	-	a	VT Transplant
-	-	b	VT/P Transplant
Full-band	-	c	VT Conversion
DWT	-	d	VT Conversion
Sel. Pre-emp.	-	e	VT Conversion
Full-band	Mean-Var.	f	VT/P Conversion
Full-band	Segmental	g	VT/P Conversion
DWT	Mean-Var.	h	VT/P Conversion
DWT	Segmental	i	VT/P Conversion
Sel. Pre-emp.	Mean-Var.	j	VT/P Conversion
Sel. Pre-emp.	Segmental	k	VT/P Conversion

Table 2: Voice conversion methods tested.  
(VT: Vocal Tract, P: Pitch)

Ten subjects have listened to 112 triples of sound files. The first and the second file were the original recordings of the source and target speakers. The third utterance contained the output to be evaluated by the subject. It was one of the following: an original recording, a transplantation output (rows 1-2 of Table 2), a conversion output (rows 3-11 of Table 2). The subjects have assigned three scores: Identity, Confidence, and Quality. The identity score is obtained by mapping the following decisions on a numerical scale: “Source”, “In between”, and “Target”. These decisions are mapped to 0.0, 0.5, and 1.0 respectively. The subject have also assigned confidence scores on their identity decisions and quality scores for the quality of the output. Both scores were in the range 1-5, 1 corresponding to low confidence (quality) and 5 corresponding to high confidence (quality). The subjects were told to choose “In between” and assign the lowest confidence score when the output sounded like a third speaker. All scores were normalized to unity and the mean and the inter-quartile ranges (IQRs) are calculated. The mean scores are shown in Fig. 7. Each group of lines in Fig. 7 correspond to another combination of the gender of the source and the target speakers. As an example, M→F is the case when the source is a male speaker and target is a female speaker. “Overall” corresponds to the case when the scores are calculated for all triples disregarding the gender information. Note that there are 14 lines in each gender combination corresponding to the output types in Table 2.

In Fig. 7, we observe that converting only the vocal tract does not produce convincing results. Even the vocal tract transplantation case was evaluated as in between the source and the target speaker. Vocal tract conversions generally had higher identity scores when the source is a male speaker. All voice conversion methods produce more convincing results in terms of similarity to the target speaker when pitch transformation strategies are involved. However, the confidence and the quality scores decrease as the amount of processing increases. Vocal tract only transplantations and conversions were assigned higher scores in terms of confidence and quality.

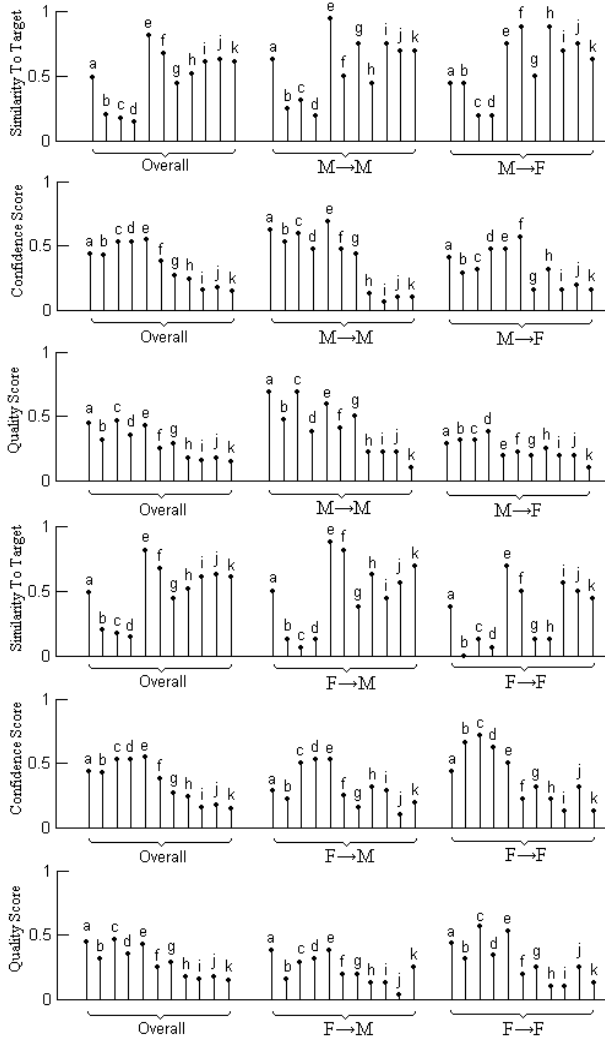


Figure 7: Subjective test results

We observe different tendencies as the gender of the source and target speaker pairs change for different vocal tract conversion methods. The full-band based method is more robust in different gender combinations. The segmental pitch model improves the identity scores. We have used IQRs as a measure of the agreement of subject decisions and for our inferences above. Low IQR is desired because it indicates that the scores are not wide spread. The performance of the full-band based method is improved when pre-emphasis was employed. The full-band based method performed better as

the results are compared with the results of the subjective test described in [7]. The source, target and third speakers were identified perfectly. We did not include the scores for these cases in Fig. 7. The identity score for the source speaker was low indicating that average similarity to target speaker was less. The target speaker had high identity scores (close to 1.0). The subjects have responded with identity scores close to 0.5 with low confidence scores in the case of “third speaker” as expected. The quality scores for all original recordings were close to 1.0.

## 5. Conclusions

In this study, we have developed two new methods for vocal tract and pitch contour transformation. The first method, selective pre-emphasis, employs band-pass filtering for detailed vocal tract spectrum estimation at a lower prediction order as compared to the full-band LP analysis. In fact, it is similar to increasing the order in full-band analysis. However, this is not generally possible for sampling rates of 44.1 KHz or higher which are especially used in dubbing applications. It is possible to increase the spectral resolution by employing more sub-bands at a constant prediction order using selective pre-emphasis. Another advantage is the possibility to employ variable prediction orders at different sub-bands providing great flexibility in the voice conversion algorithm design. We have also developed a segmental pitch contour model for more detailed pitch contour transformation. We have evaluated the performance of the new methods in a subjective test by comparing them with existing ones. This subjective test provides a useful framework for comparison of different voice conversion methods.

Although we have demonstrated the applications of the new methods in voice conversion, it is possible to propose other applications as these methods provide useful models of the vocal tract spectrum and the pitch contour. As an example, the segmental pitch contour model can be used to generate pitch contours for synthesis applications.

## 6. References

- [1] Turk, O., *New Methods For Voice Conversion*, M.S. Thesis, Bogazici University, 2003.
- [2] Gutierrez-Arriola, J.M., Hsiao, Y.S., Montero, J.M., Pardo, J.M., and Childers, D.G., “Voice Conversion Based On Parameter Transformation”, *Proc. of the ICSLP 1998*, Vol. 3, pp. 987-990, Sydney, Australia.
- [3] Stylianou, Y., Cappe, O., and Moulines, E., “Continuous Probabilistic Transform for Voice Conversion”, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, 1998, pp. 131-142.
- [4] Arslan, L.M., “Speaker Transformation Algorithm Using Segmental Codebooks”, *Speech Communication* 28 (1999), pp. 211-226.
- [5] Kain, A.B., and Macon, M., “Personalizing A Speech Synthesizer by Voice Adaptation”, in *Proc. of the 3<sup>rd</sup> ESCA/COCOSDA International Speech Synthesis Workshop*, 1998, pp. 225-230.
- [6] Chappell, D.T., and Hansen, J.H.L., “Speaker-Specific Pitch Contour Modeling and Modification”, in *Proc. of the ICASSP 1998*, Vol. II, pp. 885-888, Seattle, USA.
- [7] Turk, O., and Arslan, L.M., “Subband Based Voice Conversion”, in *Proc. of the ICSLP 2002*, Vol. 1, pp.289-292, Denver, Colorado, USA.