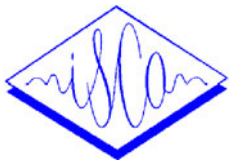


7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

On Transforming Spectral Peaks in Voice Conversion

Elizabeth Godoy [1], *Olivier Rosec*[1], *Thierry Chonavel*[2]

[1] Orange Labs R&D TECH/ASAP/VOICE, Lannion, France
[2] Télécom Bretagne, Signal & Communications Department, Brest, France
{elizabeth.godoy,olivier.rosec}@orange-ftgroup.com, thierry.chonavel@telecom-bretagne.eu

Abstract

This paper explores the benefits of transforming spectral peaks in voice conversion. First, in examining classic GMM-based transformation with cepstral coefficients, we show that the lack of transformed data variance ("over-smoothing") can be related to the choice of spectral parameterization. Consequently, we propose an alternative parameterization using spectral peaks. The peaks are transformed using HMMs with Gaussian state distributions. Two learning variants and post-processing treating peak evolution in time are also examined. In comparing the different transformation approaches, spectral peaks are shown to offer higher inter-speaker feature correlation and yield higher transformed data variance than their cepstral coefficient counterparts.

Index Terms: voice conversion, spectral transformation, spectral peaks

1. Introduction

Spectral transformation plays a crucial role in Voice Conversion (VC), both in identifying speakers' voices and ensuring high quality synthesis. The goal of spectral transformation is to transform the spectral envelope of a (source) speaker into that of a different (target) speaker. The transformation methodology can be described in three stages: first, analysis of the speech signal in order to extract spectral envelope parameters; second, training through learning a mapping between the source and target parameters; third, transformation of the source parameters to estimate those of the target. Based on this methodology, the performance of a VC system depends on two key factors: i) the choice of spectral parameters and ii) the choice of model for learning and transformation.

Traditional approaches to spectral transformation typically use Gaussian Mixture Models (GMM) [1] on cepstral coefficients or Line Spectral Frequencies (LSF). These approaches generally succeed in capturing and reproducing certain characteristic traits of the target speaker. However, the transformed data in these cases exhibits little variance, a problem often called "over-smoothing," [2], [3]. Chen et al. showed in [2] that this lack of variance in the transformed data results from a weak correlation between the source and target parameters. In addressing this problem, Chen et al. assume that the target variance is the same as that of the source and suggest a MAP adaptation algorithm to adjust the transformation function. Alternatively, in [3], Toda et al address this problem by also modifying the transformation function, but with the introduction of a "global variance" parameter to ensure that the transformed data variance mimics the target variance. In both of these cases, the "over-smoothing" problem is attributed to the transformation model and heuristics are introduced in order to increase the transformed data variance.

Fundamentally, the small transformed data variance is a result of low correlation between the source and target spectral features, as captured in the transformation model. There exist two possible explanations for this low inter-speaker feature correlation. First, this problem could be attributed to the transformation model, as in the previously mentioned works. Explicitly, the "mixing" of the data may destroy inherent inter-speaker correlation. This erroneous mixing can translate into a source-to-target mapping problem, commonly referred to as the "one-to-many" problem, [4]. The second possible explanation for the low inter-speaker feature correlation could be that the chosen spectral parameters are not capturing a meaningful link between the source and target speech. While the first hypothesis has often been assumed in related works, this paper seeks to address the second. Specifically, we can alleviate the "one-to-many" mapping problem by following the work in [4] and introducing context-dependent parameters into the GMM modeling, creating a "Phonetic GMM." In using a Phonetic-GMM, we then effectively reduce the problems resulting from the transformation model choice and can consequently focus our problem analysis on the transformation parameter choice.

In this paper, we will show that, even when ensuring correct mappings between the source and target features (on a phoneme-level), there still remains a low inter-speaker feature correlation in a classic transformation approach. Explicitly, these results indicate that the problem of low-correlation between the source and target features is due largely to the parameter choice (in this particular case, the cepstral coefficients) rather than the choice of transformation model. Consequently, we seek an alternative spectral parameterization that can better capture a meaningful link between the source and target speech. Specifically, we examine the use of spectral peaks as an alternate parameterization for voice conversion.

The structure of this paper is as follows. Section 2 begins by defining some general notation and metrics for transformation evaluation. These metrics are then applied to a classic approach to VC using discrete cepstral coefficients (DCC) in a Phonetic GMM, "DCC-GMMP." This evaluation shows that the chosen parameters, as expressed in the model, exhibit low inter-speaker correlation and are thus inadequate for conversion. In section 3, an alternative parameterization for the spectral envelope, along with an adapted model for transformation, is presented. Specifically, we consider spectral peaks and their transformation using a Hidden Markov Model (HMM) with Gaussian-state distributions, the "Peak-HMM." Two variants in the model learning related to the alignment between the source and target models are also described. In section 4, the different approaches, DCC-GMMP and Peak-HMM (with variants), are compared using a common reference for the spectral envelope. Additionally, in section 5, a post-processing technique that treats spectral peak evolution in time is examined. In section 6, a subjective evaluation of the transformation results based on informal listening tests is

discussed. Finally, in section 7, we conclude our evaluation and discuss avenues for future work.

2. Spectral Transformation Evaluation

Before considering the metrics for evaluating spectral transformation, we begin by introducing some general notation. Let us consider $n = 1,\ldots,N$ aligned source and target speech frames. In order to avoid erroneous inter-phoneme source-to-target mappings, these frames are classified by phoneme into Q model classes, i.e. each model class q represents a particular phoneme, as in [4]. Let x and y represent a single dimension of the source and target spectral parameterization, respectively. In this work, we consider that each spectral parameter dimension is independent, corresponding to a constraint that all covariance matrices be diagonal. For each class q, we consider the sample mean $\mu_q(p)$, sample variance $(\sigma_q(p))^2$ and sample cross-covariance $(\sigma_q^{xy}(p))^2$ of the p^{th} component of the spectral envelope representation of total order P. Assuming a Gaussian distribution for each dimension of the source and target spectral parameters, the transformation function for $x_n(p)$ is the Maximum Likelihood (ML) Estimator, $\hat{y}_n(p)$, given by

$$\hat{y}_n(p) = \mu_q^y(p) + \left(\frac{\sigma_q^{xy}(p)}{\sigma_q^x(p)}\right)^2 \left(x_n(p) - \mu_q^x(p)\right). \quad (1)$$

All of the transformation functions considered in this work follow (1), as is the case in many works on spectral transformation, notably [1]-[4], which use the form of this ML estimator in transformation but with variants depending on the form of the covariance matrices or using weighted mixtures of this function. With the notation defined above, we can now consider formalizing evaluation metrics.

2.1. Metrics for Evaluation

In this paper, we will consider three criteria for the evaluation of spectral transformation. First is the strength of the link between the source and target parameters in the model. Formally, this is expressed in the correlation. Specifically, we consider the average correlation between source and target parameters in the model

$$\rho^{XY} = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{1}{P}\sum_{p=1}^{P}\left|\frac{\left(\sigma_q^{xy}(p)\right)^2}{\sigma_q^x(p)\sigma_q^y(p)}\right|\right). \quad (2)$$

This criterion is critical in determining the capacity of the parameters in the model for transformation, as the source-target feature correlation scales the factor in (1) that is dependent on the source data to be transformed. Similarly, the variance of the transformed data will depend on this factor and, thus, the correlation. Generally, the variance of the transformed data captures the influence of the correlation in the transformation results. Accordingly, the second criterion that we consider compares the transformed variances for each class to those of the target. Specifically, we consider the average ratio of the variances, *VR*,

$$VR = \sum_{q=1}^{Q}\frac{N_q}{N}\left(\frac{1}{P}\sum_{p=1}^{P}\left(\frac{\sigma_q^y(p)}{\sigma_q^y(p)}\right)^2\right), \quad (3)$$

where $\sigma_q^y(p)$ represents the sample variance of the transformed data and N_q frames are considered in class q. Finally, for an indicator of the transformation quality, we consider the absolute error between the transformed and target frame envelopes; specifically, the Mean Squared Error (MSE) normalized by the target parameter energy:

$$\varepsilon = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{P}\sum_{p=1}^{P}\left(\hat{y}_n(p) - y_n(p)\right)^2\right)}{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{P}\sum_{p=1}^{P}\left(y_n(p)\right)^2\right)}. \quad (4)$$

Together, these three criteria (2)-(4) form a complete evaluation of spectral transformation, both of the approach and the results.

2.2. Speech Data

Our speech data is taken from corpora used in France Télécom's speech synthesis system *Baratinoo*, which contains speech sampled at 16kHz whose phonetic labeling and segmentation is manually verified. Currently, we consider transforming only vowels, as these are among the most important phonemes in speaker identification. In this work, a parallel corpus consisting of a female (source) and male (target) speaker is used. The source and target speech frames are analyzed pitch synchronously. The three center ("stable") frames of each source and target phoneme are automatically aligned. The remaining frames are aligned uniformly in time, within each phoneme. The training and test data sets each consist of 100 distinct phrases (roughly 30,000 aligned frames per set).

2.3. Evaluating "Classic" Spectral Transformation

Given the evaluation criteria described in section 2.1, we can now re-visit a classic approach to spectral transformation. In particular, we consider DCCs, as described in [1], with no cutoff frequency and no frequency-scale warping. In this case of spectral transformation using DCCs, P represents the cepstral order: thus, $x_n(p)$ is the p^{th} cepstral coefficient of the n^{th} source frame. Examining the evaluation metrics described in section 2.1 yields the results summarized in Table 1. We consider the correlation for different cepstral orders in parentheses; since higher order coefficients capture more detail, we can expect less correlation as we increase the cepstral order. Additionally, we have included two MSE scores: ε as defined by (1) and (4) and ε_{mean}, which is the MSE of transformed data calculated using only the target mean in (1), corresponding to a VQ-type conversion scheme.

Table 1. *Evaluation Results: Classic Transformation*

ρ^{XY} order 40 (20, 10)	0.08 (0.12, 0.16)
VR	0.02
ε	-8.46 dB
ε_{mean}	-8.19 dB

The results in Table 1 show weak links between the source and target parameters, as evident by the low average correlation (for all examined cepstral orders). Accordingly, the low ratio of variances shows that there is very little variation in the transformed data. What's more, the difference between the MSE using the entire transformation function versus only the mean is a fraction of a dB. Hence, the estimated target parameters are essentially the target means. These results verify those in [2] and [3]. However, in this case, one-to-one mappings between the source and target frames, within a phoneme, are ensured. Based on these observations, we hypothesize that the lack of inter-speaker feature correlation is primarily due to the parameter choice. Consequently, we seek an alternative spectral parameterization for transformation, namely spectral peaks.

3. Transforming Spectral Peaks

3.1. Peak Representation & Analysis

Similarly to [5]-[6], we model the spectral envelope for frame n as a sum of Gaussian peaks

$$E_n(f) = \sum_{m=1}^{M_n} a_n^m \exp\left(-\frac{(f - f_n^m)^2}{2v_n^m}\right), \qquad (5)$$

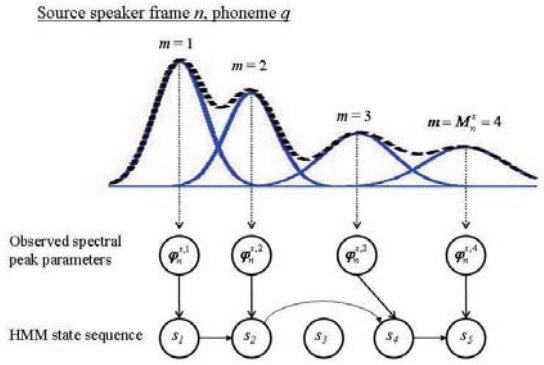
where f indicates frequency and M_n is the number of peaks in frame n. The number of peaks for each frame is not fixed but is limited to 20. The parameters $\varphi_n^m = \left[f_n^m, a_n^m, v_n^m\right]^T$ represent the frequency, amplitude and variance of the m^{th} peak in frame n. As discussed in [5] and [6], this representation offers an intuitive and flexible representation for the spectral envelope in a conversion context, while maintaining a good analysis-synthesis speech quality.

For the peak analysis, as in [6], the Gaussian peak parameters are selected from peak-picking directly on the Discrete Fourier Transform (DFT), using a frequency mask to avoid modeling harmonic peaks and to increase resolution in regions more sensitive to human hearing. The peak variance is then calculated to fill-in the envelope in-between peak amplitudes. We note here that, given this parameter estimation, the spectral peak variance does not carry a physical meaning. Consequently, later in learning, this parameter is not considered in determining model classes. Finally, for the current work, we do not use the inter-frame alignment described in [6], as we do not currently consider the evolution of spectral parameters in time in the analysis stage of VC.

3.2. Peak Sequence Modeling

The number of peaks determined from the analysis described above can vary for each source and target frame. Thus, there is no inherent intra or inter-speaker alignment between peaks and the Phonetic GMM described in section 2.2 cannot be directly applied. In order to model the source and target speaker spaces with this peak representation, we consider the spectral envelope as a sequence of peaks in frequency. Explicitly, for frame n of the source speaker, we have the following sequence $X_n = \left[\varphi_n^{x,1}, ..., \varphi_n^{x,m}, ..., \varphi_n^{x,M_n^x}\right]$ of spectral peak parameters. The ensemble of source (or target) peak sequences, for a particular phoneme, can then be modeled by an HMM, as in [7]. We consider here left-right HMMs in which skips between states are allowed. We refer to this modeling of spectral peaks using a single HMM per phoneme as a Peak-HMM. A diagram illustrating the peak sequence modeling is shown in Figure 1.

Figure 1: Illustration of HMM Peak-Sequence Modeling for a source speaker frame containing M_n=4 spectral peaks using a HMM with P=5 states



3.3. Peak-HMM Learning

The following section describes the Peak-HMM learning: first, learning of a single HMM for each phoneme (for each speaker individually); second, learning a mapping between the source and target HMMs for each individual phoneme. The overall learning procedure is summarized as follows.

Peak-HMM Learning: (For Phoneme q)

Data: $X_n, Y_n : \forall n \in N_q$

I. Estimate Source & Target HMMs (independently)
  i. Data clustering: Generate Gaussian Classes (States)
  ii. Calculate HMM transition probabilities
II. Joint Source-Target Space
  i. Inter-Speaker State Alignment (2 Proposed Methods)
  ii. Calculate Cross-Covariance

First, all of the source (or target) peaks (frequency & amplitude) are grouped using a simplified GMM with 20 classes (I. *i*). The simplification consists in using a MAP constraint on the EM algorithm so that each peak is associated with a single class. Statistically insignificant classes are then removed. These Gaussian classes then form the states in the phoneme HMM. In the second step (I. *ii*), the transition probabilities and initial probability distribution for the speaker HMMs are calculated. Unlike [7], we do not currently consider the time dimension.

Second, in order to learn mappings between the source and target HMMs for a given phoneme, an alignment between the source and target HMM states in a phoneme must be

determined (II. *ii*). Two methods for this inter-speaker state alignment will be described in the following subsections.

Given the particular inter-speaker state alignment, in the final step (II. *ii*), the sample cross-covariance $(\sigma_q^{xy}(p))^2$ for each source-target state pair must be calculated in order to apply (1) to transform each source peak parameter $x_n(p)$. This calculation considers source and target peaks, from the same time-aligned joint frame, that are realizations of the states (classes) in the particular source-target state pair. More explicitly, consider that the p^{th} source and target states are aligned. Let $I_q^x(p)$ and $I_q^y(p)$ denote the set of indices of source and target speech frames, respectively, that contain a peak realization belonging to the HMM state p of phoneme q and let $I_q^{xy}(p) = I_q^x(p) \cap I_q^y(p)$. Then, the cross-covariance is calculated as follows

$$\left(\sigma_q^{xy}(p)\right)^2 = \frac{1}{\left|I_q^{xy}(p)\right|} \sum_{n \in I_q^{xy}(p)} \left(x_n(p) - \mu_q^x(p)\right)\left(y_n(p) - \mu_q^y(p)\right) \tag{6}$$

where |.| denotes the cardinal of a set. Note that for the Peak-HMM, $x_n(p)$ represents one of the peak parameters (frequency, amplitude, or variance) for the p^{th} HMM state.

3.3.1. Method 1: One-to-One Sequence Alignment

The first method for inter-speaker state alignment simply imposes a one-to-one alignment between the source and target classes. That is, the alignment follows the ordering in frequency of the classes (i.e. state p of the source HMM is aligned to state p of the target HMM). In cases for which the number of source and target states is not identical, the final source states (highest in frequency) are repeated or removed in order to match the number of target HMM states. With the possible exception of source classes representing the highest peak frequencies, this alignment ensures both full representation of the source and target classes as well as one-to-one source-to-target mappings (i.e. no repeated classes in the source or target spaces).

3.3.2. Method 2: Aligning Most Probable Peak Sequences

The second proposed method involves comparing the class mean frequencies. Considering only the HMM statistics, we estimate the most-likely state sequence for the target and source speaker. Each of the most-likely target states is aligned to the most-likely source state closest in frequency. Each remaining un-aligned source state (most likely or not) is then aligned to the target state (most likely or not) nearest in frequency. With this method, we assure coverage of the most probable target classes and all of the source classes while limiting "warping" of the frequency axis in transformation by aligning source-target classes nearest in frequency.

3.4. Transformation

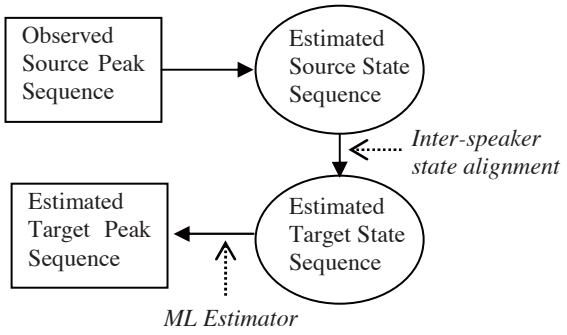
The following diagram in Figure 2 describes the transformation process, namely how to transform an observed source peak sequence into an estimated target peak sequence.

The first step in this transformation is to find the most likely source state sequence given the observed source peak sequence. This problem can be solved using a Viterbi algorithm, as described in [8]. Each state in this source state sequence corresponds to a target state, according to the inter-speaker state alignment determined in the Peak-HMM learning.

Given these target states estimated from the inter-speaker state alignment, the most probable target state sequence is determined, with an addition of target states, if necessary. Specifically, the target states estimated from the inter-speaker state alignment serve as "anchors" in calculating the estimated target state sequence; that is, each of these states is fixed as a member of the estimated target state sequence. Additional target states can be introduced into the estimated state sequence by the following process: given the i^{th} state in the estimated target state sequence, the most-likely transition, to a $(i+1)^{th}$ state, is considered; if this $(i+1)^{th}$ state is not one of the anchors and if this transition does not exclude an anchor, then this state is added to the estimated target state sequence.

In the final estimation step of the Peak-HMM transformation, for each target state related to an observed source peak, the ML estimator (1) is used to estimate the corresponding target peak parameters. The estimated target peaks from the remaining target states in the sequence, if any, are taken as the target state mean. Finally, given the estimated target peak sequence, the estimated target envelope is generated from (5).

Figure 2: Peak-HMM Transformation



4. Evaluation Results

As in the case evaluating a classic spectral transformation in section 2.2, the capability of a chosen model to transform the chosen spectral parameters is indicated by the average correlation (2). In the case of the Peak-HMM, we consider the average correlation of the peak frequency, amplitude and variance, respectively. Examining the average correlation for each peak parameter using (6) in (2), considering each variant in the learning, we have the following results shown in Table 2. For each proposed method in the Peak-HMM, the correlation between peak parameters is nearly identical. This similarity shows that the source-target inter-speaker feature correlation does not depend significantly on the chosen source-to-target alignment. This observation indicates that there exists an overall inter-speaker correlation between the ensembles of peaks that is not a direct result of the model constraints on source-to-target mappings.

Of the three parameters in Table 2, the peak log-amplitude is the most relevant. Considering the peak frequency, transformation of this parameter is essentially carried out in

selecting the state sequence. Significant variations in frequency will not exist within the model states, as this would correspond to a change in state. Considering the peak variance, as previously discussed in section 3.1, this is a less important parameter in transformation. Consequently, the most significant indication of the Peak-HMM's capability for transformation is given by the average correlation of the peak amplitude (log-amplitude).

Comparing the correlation values in Table 2 with those for the classic transformation approach in Table 1, we find a significant increase in correlation using spectral peaks rather than DCCs, especially for the log-amplitude. In other words, the link between the source and target parameters, as expressed in the model, is stronger in the Peak-HMM. It must be noted that the parameter spaces, and consequently model dimensions, are different in Table 1 and Table 2. However, the average correlation shows the strength of the link between these different spectral parameters, as expressed in their respective models and as used in transformation with (1).

Table 2. *Peak-HMM Parameter Correlation*

parameter	**correlation**	
	Method (1)	Method (2)
frequency	0.10	0.10
log(amplitude)	0.39	0.38
sqrt(variance)	0.27	0.26

In order to examine the accuracy of the Peak-HMM in estimating the target parameters, we need to consider the remaining evaluation metrics in (3) and (4). Additionally, we seek to compare the Peak-HMM results with those of the more classical approach to transformation described in section 2. In order to compare the transformed envelopes of both methods on equal footing: i) a common reference for both approaches must be considered and ii) envelopes must be considered in the same domain. For the common reference envelopes for the source and target speakers, we select the peak envelope calculated from the DFT, given by (5). These reference envelopes are parameterized and transformed using the different approaches (GMM-DCCP and Peak-HMM). The resulting transformed enveloped are then all evaluated in the discrete cepstral domain. More specifically, for the phonetic GMM, the DCCs (order 40) are calculated from the reference peak envelopes and the corresponding model and results are examined in the discrete cepstral domain. Note that the reference envelopes are not the same as in section 2.2, thus, the results could change from Table 1. However, we state here that parameter correlation for cepstral order 40 remained the same as in Table 1, 0.08. For the Peak-HMM (including all variants), learning and transformation are carried out as described in section 3 in the spectral peak domain. The resulting transformed envelopes are then parameterized with DCCs (order 40). In this parameterization, we consider frequencies up to the final peak, as the drop-off past this peak can be significant, thus influencing the resulting DCCs. Applying the metrics (3) and (4) to both transformation results, all in the discrete cepstral domain, we have the following results in Table 3.

Table 3. *Evaluation Results: DCC-GMMP vs Peak-HMM*

	DCC-GMMP	Peak-HMM Method (1)	Peak-HMM Method (2)
VR	0.01	0.41	0.34
MSE: ε	-7.86 dB	-5.29 dB	-5.06 dB

In Table 3, there is significantly larger similarity between the transformed and target data variance for the Peak-HMM as compared to the Phonetic GMM with DCCs. Note that, unlike the work in [2] and [3], this variance is not a result of heuristic constraints introduced in the transformation function, but rather a result of the differences in the transformation domain; notably, a difference in parameter choice and, consequently transformation model. Considering the *MSE*, we see that the DCC-GMMP gives higher accuracy in a frame-by-frame transformed-target comparison. This result can be expected as GMM-based transformation is intended to minimize the overall mean squared error in the discrete cepstrum domain. Among the Peak-HMM variants, the method for inter-speaker state alignment with one-to-one mappings outperforms the other variant. The frame-by-frame envelope comparison indicates that the Peak-HMM (in all cases) is currently lacking in estimation accuracy, according to the objective metrics examined here. We will later discuss observations on informal subjective evaluations of the different transformation approaches. Nonetheless, the stronger source and target links for the Peak-HMM and the ability to better capture the variation in the target spectral envelope show that this type of approach holds promise for spectral transformation.

5. Post-Processing of Spectral Envelope Discontinuities

Discontinuities in the transformed spectral envelope between adjacent frames can generate artifacts that diminish the transformed speech quality, as described in [2] and [5]. In [2], median and lowpass filtering are employed to smooth discontinuities in a sequence of transformed envelopes. Alternatively, the work in [5] considers "event functions" to smooth the evolution of spectral peaks across a sequence of frames. In this work, we propose a type of median filtering of transformed spectral peak parameters across a sequence of frames within a phoneme. Specifically, beginning with the center frame of the phoneme, we average the transformed peak parameters with those of the frames immediately to the left and to the right. The peaks between two frames are aligned by locally minimizing the distance between peak locations in frequency, as proposed in the analysis stage in [6]. Peaks of the center frame that are not aligned with peaks from the neighboring frames on either side are removed. This process of aligning peaks in frequency and averaging the transformed peak parameters is continued for each frame individually, moving outward from the center frame (to the left and to the right) to the phoneme boundary. Applying this post-processing technique to the frames transformed using the Peak-HMM (for all alignment methods), we have the following results shown in Table 4.

Table 4. *Post-Processing Evaluation Results: Peak-HMM*

	Method (1)	Method (2)
VR	0.36	0.28
MSE: ε	-5.39 dB	-5.17 dB

These results show that the averaging of transformed spectral peak parameters in time reduces the transformed data variance by approximately 0.05. However, the MSE is improved by about 0.1dB in both cases. While these objective results do not show a significant difference using the post-processing considering the evolution of spectral peaks in time, the following section notes an important improvement in subjective quality.

6. Subjective Evaluation: Informal Listening Tests

Informal listening tests were conducted on a selection of phrases in order to compare the converted speech quality. Specifically, an HNM [1] is used in analysis and synthesis. In order to evaluate the transformed spectral envelope, we consider the target speech, with only the harmonic amplitudes converted. That is, the harmonic amplitudes from the original target speech analysis are replaced with harmonic amplitudes sampled from the transformed envelopes. In this way, we are able to isolate the effect of only the spectral envelope on the converted speech quality.

First, we note that, while the absence of a peak in the spectral envelope (especially in mid-to-high frequencies) may significantly increase the MSE in a frame, this absence might not significantly affect the perceptual quality. This indicates that, while the MSE may be high, the perceptual quality is not necessarily poor. This is particularly relevant in the case of transforming peaks because errors are often localized in certain regions of the frequency spectrum, unlike the case of transforming cepstral coefficients in which errors in transformation affect all frequencies.

Second, in comparing the different alignment methods for the Peak-HMM, neither could be consistently judged as superior to the other.

Third, in all examined cases for the Peak-HMM, the post-processing alignment never worsened the quality. For the sections of speech exhibiting high quality, no degradation was perceived. For artifacts resulting from spectral discontinuities between frames, the proposed post-processing improved the perceived quality in the majority of cases. These observations seem to follow those found in [2], though more thorough evaluation should be carried out to confirm this.

Finally, in comparing informal observations on the overall quality of the GMM-based transformation of the cepstral coefficients versus HMM-based transformation of spectral peaks, we make the following initial remarks. In the GMMP-DCC case, we note a "muffling" or "loss-of-presence," a degradation that is always perceived continuously. Conversely, it seems that the Peak-HMM can, in some instances, yield a higher converted speech quality comparable to the analysis-synthesis quality. However, in other cases, the converted speech can also be severely degraded and sound very unnatural. These degradations could result from both problems in the source-to-target state mappings and in the peak classification. Overall, these observations indicate that there is potential in using a Peak-HMM approach in VC in order to achieve high converted speech quality.

In this section, the authors have given informal observations on the converted speech quality. It must be emphasized that the original prosody and harmonic phases from the analysis of the target speech have been kept in all cases; the only feature that we are evaluating is the spectral envelope conversion. More conclusive results should involve more formal testing and conversion with a wider variety of speakers.

7. Conclusions & Future Work

This work has shown that aspects of the "over-smoothing" problem in spectral transformation can be reduced by choosing an adequate spectral parameterization. Spectral peaks have been shown to better capture the correlation between source and target speech, as compared to cepstral coefficients. While the transformation accuracy needs to be improved, the increased inter-speaker feature correlation and, consequently, the increase in transformed data variance, demonstrate promise in using spectral peaks for voice conversion.

Further work will be conducted in order to find a more robust peak classification and source-to-target state mapping. One possible approach for this could be to incorporate the time-evolution of spectral peaks in analysis, as in [6], and in learning the transformation model, as in [7].

Furthermore, more extensive subjective testing and evaluation with a wider variety of speakers will be necessary to examine the different methods for transformation more thoroughly.

8. References

[1] Stylianou, Y. "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. diss., ENST, Paris, France, Jan. 1996.

[2] Chen, Y., Chu, M., Chang, E., and Liu, J., "Voice conversion with smoothed GMM and MAP adaptation", in Proc. of Eurospeech '03, pp 2413-2416.

[3] Toda, T., Black, A., and Tokuda, K., "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in Proc. of ICASSP '04, Vol. 1. pp. 9-12.

[4] Godoy, E., Rosec, O., and Chonavel, T., "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modeling", in Proc. of Interspeech '09, pp. 1627-1630.

[5] Nguyen, B. and Akagi, M., "Spectral modification for voice gender conversion using temporal decomposition," Journal of Signal Processing, Vol. 11, No. 4, pp. 333-336, July 2007.

[6] Godoy, E., Rosec, O., and Chonavel, T., "Speech spectral envelope estimation through explicit control of peak evolution in time", in Proc. of ISSPA '10.

[7] Rentzos, D., Vaseghi, S., Yan, Q., and Ching-Hsiang, H., "Voice conversion through transformation of spectral and intonation features," in Proc. of ICASSP '04, Vol 1, pp. 21-24.

[8] Rabiner, L.R., "A tutorial on hidden markov models and selected applications in speech recognition," in Proc. of the IEEE, Vol. 77, No 2, February 1989, pp. 257-286.