Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling

Elizabeth Godoy, Olivier Rosec, Thierry Chonavel

Alleviating the One-to-Many Mapping Problem in Voice Conversion with Context-Dependent Modeling

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

[1] Orange Labs, Lannion, France
[2] Telecom Bretagne, Signal & Communication Department, Brest, France
{elizabeth.godoy, olivier.rosec}@orange-ftgroup.com, thierry.chonavel@enst-bretagne.fr

Abstract

This paper addresses the "one-to-many" mapping problem in Voice Conversion (VC) by exploring source-to-target mappings in GMM-based spectral transformation. Specifically, we examine differences using source-only versus joint source/target information in the classification stage of transformation, effectively illustrating a "one-to-many effect" in the traditional acoustically-based GMM. We propose combating this effect by using phonetic information in the GMM learning and classification. We then show the success of our proposed context-dependent modeling with transformation results using an objective error criterion. Finally, we discuss implications of our work in adapting current approaches to VC.

Index Terms: Voice conversion, GMM, Spectral mapping.

1. Introduction

Voice Conversion (VC) is the process of modifying the speech of a (source) speaker so that it sounds as if a different (target) speaker uttered the same phrase. VC finds applications in many systems involving speech, from messaging and translation systems to film dubbing. One particularly relevant application is for text-to-speech (TTS) systems in which VC can be used to create different voices without need for recording an entirely new database. While VC is pertinent to many applications, the subject is also highly complex and there remains much work to be done in improving the quality of current VC technologies.

The classic approach to VC involves extraction and modification of acoustic parameters in speech, mainly spectral envelope and pitch, which have been shown to be among the most relevant parameters for speaker identification [1],[2]. Following this methodology, the first step in VC is to align speech frames from two speakers, in time, so that they correspond to roughly the same events. The next step is to describe and learn a mapping between source and target spaces. After this learning, the first step in transformation is classification (decoding) of the source frames according to the learned source space. Finally, a transformation function is applied directly on acoustic parameters of the source speech in order to estimate the target speech. There are many different methods used to describe the source and target spaces and their corresponding mapping, including Vector Quantization (VQ) [3], Neural Networks (NN) [4], and Gaussian Mixture Models (GMMs) [1],[5]. GMMs are the most employed in VC as they have been shown to outperform the other methods [6]. Following the general approach to VC above, our work focuses on spectral transformation using a GMM.

It is important to note that the VC methods described above depend heavily on time-alignment of the source and target speech frames. Using parallel speech corpora (e.g. identical spoken texts), time alignment can be constrained to ensure the same phonetic contexts for each joint frame. However, even this phonetically-constrained time alignment does not ensure an acoustic alignment. That is, acoustically dissimilar events, mainly frames corresponding to different vocal tract configurations for each speaker, could be aligned (e.g. one speaker could have markedly different ways of saying 'A,' though these instances will always be aligned to a singular 'A' from the other speaker). After joint source/target frames are determined, grouping of these frames in the learning stage is acoustically-based. Accordingly, the lack of acoustic alignment between the source and target parameters in the joint frame gives rise to the widely recognized "one-to-many" problem [7]. More explicitly, the one-to-many problem occurs when a single group of source frames with similar spectral characteristics is associated with target frames in multiple groups representing different acoustic instances. In this case, it is impossible for the distinct target events to be uncovered from only the source information.

Previous work has examined this problem in the context of VQ-based conversion, using hard-classification and clustering source frames independently from target frames [7], unlike our work, which considers mixtures of components and learning with joint source/target frames in a GMM-context. Moreover, to our knowledge, no solution to combat the one-to-many problem has been proposed.

In this paper, we propose taking into account contextual information (namely phonetic context) in GMM-based VC in order to alleviate the one-to-many mapping problem. We further show that this strategy greatly improves the transformation results from a traditional acoustically-based GMM. The structure of this paper is as follows. Section 2 presents the framework for GMM-based spectral conversion. In section 3, we exploit the difference between classification with source versus joint frames in order to illustrate the one-to-many mapping problem. In section 4, we introduce phonetic information into the GMM framework in order to improve classification of speech frames, thus helping to alleviate the one-to-many mapping problem. Finally, sections 5 and 6 conclude our discussion and mention avenues for future work.

2. GMM-Based VC

The following section briefly summarizes spectral conversion using a GMM, as presented in [1]. We will later refer to this model as an "Acoustic GMM." Let $X = \{x_1,...,x_N\}$ and $Y = \{y_1,...,y_N\}$ be sets of spectral feature vectors for N time-aligned frames from the source and target speech, respectively, and let $Z = (X,Y)$ be the set of joint source/target spectral vectors. A GMM represents the probability distribution of these vectors as a mixture of Q multivariate Gaussians,

$$p(z) = \sum_{q=1}^{Q} \alpha_q N(z;\mu_q,\Sigma_q), \sum_{q=1}^{Q} \alpha_q = 1, \alpha_q \geq 0 \ , \qquad (1)$$

where $N(z;\mu_q,\Sigma_q)$ is a normal distribution with mean μ_q and covariance Σ_q and α_q is the prior probability of the component q. Each component represents an acoustic class, hence, the GMM describes an acoustic space for the source and target speakers. The parameter set $\{\alpha_q, \mu_q, \Sigma_q, q=1\ldots Q\}$ is calculated from an expectation maximization (EM) algorithm on the set of joint source/target spectral vectors Z. Given the learned GMM parameters, the transformation function is the maximum likelihood (ML) estimator of the target vector, given the corresponding source vector. Specifically, this function takes the form of a weighted mixture of the ML estimator for each component

$$\hat{y}_i(x_i) = E\left[Y \middle| X = x_i\right] = \sum_{q=1}^{Q} w_q^x(x_i)\left[\mu_q^y - \Sigma_q^{xy}\left(\Sigma_q^{xx}\right)^{-1}(x_i - \mu_q^x)\right], \quad (2)$$

where $w_q^x(x_i)$ is the a posteriori probability that the frame x_i belongs to the acoustic class described by the component q:

$$w_q^x(x_i) = \frac{\alpha_q N(x_i;\mu_q^x,\Sigma_q^{xx})}{\sum_{l=1}^{Q}\alpha_l N(x_i;\mu_l^x,\Sigma_l^{xx})}. \quad (3)$$

We refer to calculation of these a posteriori mixture weights as the classification step in transformation.

3. Illustrating the One-to-Many Problem

3.1. Source vs. Joint Classification

The benefits of learning the GMM parameter set with a joint distribution (versus separate source and target models) are explored in [1]. However, the difference between using source versus joint data in transformation (specifically classification) has yet to be examined. Though it is not feasible in practice to have target data available in the classification step, exploring the difference between classification with source versus joint data provides insight into the types of source-to-target mappings present in spectral conversion. Accordingly, in addition to the mixture weights in (3), we will examine conversion results using the following joint mixture weights:

$$w_q^z(z_i) = \frac{\alpha_q N(z_i;\mu_q^z,\Sigma_q^{zz})}{\sum_{l=1}^{Q}\alpha_l N(z_i;\mu_l^z,\Sigma_l^{zz})}. \quad (4)$$

The corresponding joint transformation function is the same as in (2) with only the mixture weights changing.

We will now examine conversion results using classification with source and joint data. Our speech data is taken from two parallel corpora of French speakers, namely a female (source) and male (target), used in France Telecom's TTS system, *Baratinoo*. Both acoustic databases are sampled at 16kHz and segmented (with manual verification) into phones and diphones. We consider only voiced phonemes (28 in total, excluding 'P','T','K','S','CH','F') in transformation, as these phonemes carry most of the information related to speaker identity [2]. We take the middle frame of each phone along with its two adjacent frames, comprising the "stable" part of the phone. Such a choice has two advantages: first, the stable part of a phone is less influenced by adjacent phones, so it better represents an isolated acoustic event; second, this choice avoids need for a time-alignment algorithm as association between source and target frames is based on the corpus segmentation. We discuss transformation for all phone frames later in section 4.3. In total, we have approximately 82,000 feature vectors of which half are used for learning and half for testing. We calculate the discrete cepstrum coefficients (order 16) for each frame using the Harmonic plus Noise Model (HNM) as presented in [5] with a 10ms step and a maximum voicing frequency fixed to half the sampling frequency. We examine an objective error metric, namely the average MSE between the transformed and target frames, normalized by the average MSE between the source and target frames

$$MSE = \frac{\sum_{i=1}^{N}\left\|\hat{y}_i(x_i) - y_i\right\|^2}{\sum_{i=1}^{N}\left\|x_i - y_i\right\|^2}, \quad (5)$$

where $\hat{y}_i(x_i)$ can be calculated with either the source or joint mixture weights, as specified. We discuss subjective analyses of the conversion results later in section 5. Here, we focus first on objective measures to evaluate different conversion methods.

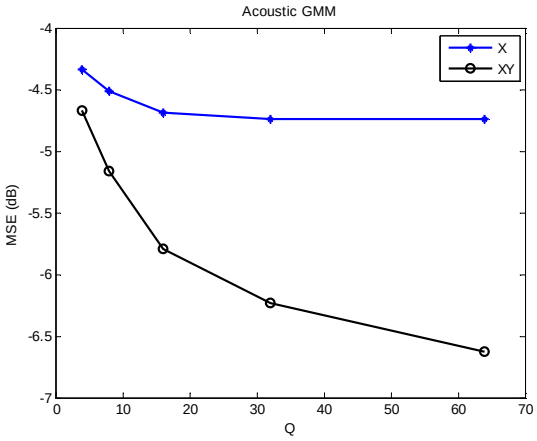

Figure 1: *MSE with Source ('X') and Joint ('XY') Decoding wrt the number of GMM components Q.*

Figure 1 plots the transformation error using source and joint mixture weights as a function of the number of GMM components Q. The two curves show that there is a significant difference in performance between using source and joint mixture weights, indicating the presence of one-to-many mappings. More explicitly, it is clear that the source-to-target mappings are not strictly one-to-one, as in the case of a one-to-one mapping, there is no difference between classification with source or joint data. In calculation of both curves, the model parameters remain unchanged; only the mixture weights differ. Consequently, in order to understand the difference in performance between classification with source and joint data, we must examine the mixture weights.

Figure 2 shows histograms of the maximum mixture weight and the corresponding MSE for joint and source classification, respectively, for the representative example Q = 16. In Figure 2a, nearly 80% of the total frames have a maximum mixture weight greater than 0.9, showing that joint classification mostly clearly favors one GMM component. On the other hand, in Figure 2b, only about 30% of the source-classified frames have a maximum mixture weight greater than 0.9, indicating that source classification distributes mixture

weight more evenly over multiple components, creating "heavier" mixtures. In both Figures 2c and 2d, we see that the lower (higher) the maximum mixture weight, the higher (lower) the MSE. These observations on source and joint classification describe a "one-to-many effect" that explains the difference in performance seen in Figure 1.

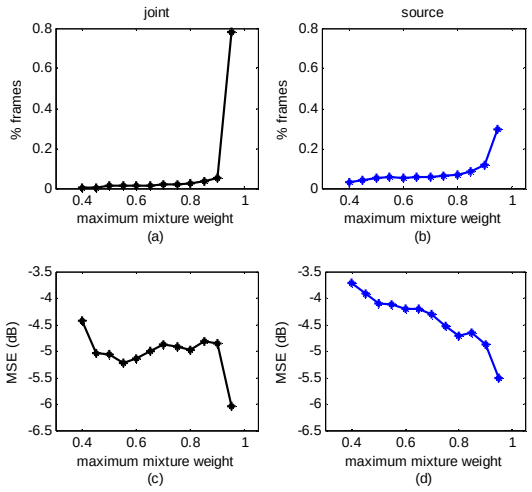

Figure 2: *Histograms of maximum mixture weight for joint (a) & source (b) classification with the corresponding MSE in (c) & (d), respectively.*

3.2. Isolating One-to-Many Mappings

We can now use the above observations to isolate frames in one-to-many mappings, according to a criterion on the source $w(x)$ and joint $w(z)$ mixture weights outlined below

$$\begin{aligned} &\text{One-to-One:} \quad w^x_{\max}(x_i) \geq \gamma \;\&\; w^z_{\max}(z_i) \geq \gamma; \\ &\text{One-to-Many:} \quad w^x_{\max}(x_i) < \gamma \;\&\; w^z_{\max}(z_i) \geq \gamma, \end{aligned} \qquad (6)$$

where γ is a threshold on the maximum mixture weight values, $w_{\max} = \max_q(w_q)$. Upon examination of a range of threshold values, we select γ = 0.8 as a representative case for this work. Results obtained from applying this mapping criterion to the test data set with Q=16 are summarized in Table 1. We see that a significant majority (more than 80%) of frames is in either a one-to-one or one-to-many mapping. The remaining frames include outliers in the data and frames in many-to-many mappings. For frames in a one-to-one mapping, there is virtually no difference in MSE between decoding with source or joint data. On the other hand, for frames in a one-to-many mapping (or numerous one-to-many mappings), there is a significant difference in MSE, of greater than 1.5dB, between decoding with source and joint frames. Thus, these results confirm that one-to-many mappings occur frequently and present a significant source of error in the classic GMM-based approach to VC.

Table 1. *MSE (dB) by mapping type, Acoustic GMM.*

GMM (Acoustic)	**One-to-One (O)**	**One-to-Many (M)**	**Remaining (R)**
% frames	47	36	17
X decoding	-5.99	-4.24	-2.97
XY decoding	-6.06	-5.75	-5.22

4. Context-Dependent Modeling

In section 3, we showed that one-to-many mappings are indistinguishable in classic GMM-based conversion, yielding "heavy" mixtures that result in high conversion error. In order to combat this problem, we propose using context-dependent parameter constraints on the GMM to improve classification by eliminating erroneous mixtures.

4.1. GMM with Phonetic Information

In particular, we introduce phonetic information into the GMM framework. Specifically, we group frames by phoneme and attribute one GMM component to each group, giving 28 in total. In learning a "Phonetic GMM", instead of an EM algorithm, Gaussian classes are generated according the phonetic label for each frame. The parameters for each Gaussian component are estimated as follows:

$$\mu_k = \frac{1}{N_k}\sum_{l=1}^{N_k} z_l, \Sigma_k = \frac{1}{N_k}\sum_{l=1}^{N_k}(z_l-\mu_l)(z_l-\mu_l)^T, \alpha_k = \frac{N_k}{N}, \quad (7)$$

where calculations are performed using the N_k feature vectors in the learning data set for a particular phoneme *k*. In transformation, we can then classify frames based on their phoneme label, avoiding mixture weight calculation. In what we refer to as transformation with phoneme separation, the transformation function takes the same form as in (2), but with w_q set to one for the component corresponding to the phoneme label of the frame. Given the model parameters of the Phonetic GMM, mixture weights can also be calculated for the source and joint feature vectors as in (3) and (4), respectively.

4.2. Conversion Results on Stable Parts

Table 2 summarizes the conversion results for the Acoustic (A) and Phonetic (P) GMMs, with each model containing 28 Gaussian components. We see that classification with phoneme separation for a Phonetic GMM greatly outperforms the classic acoustic GMM with source decoding, yielding an improvement of about 0.9dB. We also see that the Acoustic GMM performs the best with joint decoding, even though this result is not achievable in practice. Thus, with the Phonetic GMM, we make a slight sacrifice in the optimal (but not realizable) performance using joint acoustic decoding, in order to achieve significant gains using classification with phoneme separation afforded by this model.

Table 2. *Conversion Results for Phonetic & Acoustic GMM.*

GMM (Q=28)	**MSE (dB)**	**Mean($w_{\max}$)**
P,Phoneme Separation	-5.58	--
P, X Decoding	-4.71	0.72
A, X Decoding	-4.71	0.71
P, XY Decoding	-5.84	0.92
A, XY Decoding	-6.18	0.92

Moreover, we find that the Phonetic and Acoustic GMMs behave nearly identically with source decoding, both in MSE and in mixture weight distribution. These similarities suggest that the problem of one-to-many mappings exists in both models, without considering phoneme separation in classification. Accordingly, we now re-visit the problem of one-to-many mappings within the context of a Phonetic GMM.

Table 3 shows the conversion results for the Phonetic GMM with a breakdown of the MSE for one-to-one and one-to-many frames according to the criterion in (6). We see that the one-to-many effect, as discussed in section 3, is present in the Phonetic GMM with acoustic decoding. Moreover, classification with phoneme separation improves the MSE by more than 1.2dB for frames in one-to-many mappings, while the results for frames in one-to-one mappings remain virtually unchanged. Thus, learning and classification based on phoneme label is able to reduce the errors resulting from one-to-many mappings.

Table 3. *MSE (dB) by mapping type, Phonetic GMM.*

GMM (Phonetic)	**O**	**M**	**R**
% frames	35	44	21
Phoneme Separation	-5.68	-5.74	-5.07
X decoding	-5.66	-4.52	-3.85
XY decoding	-5.73	-5.99	-5.72

4.3. Results for All Phone Frames

While the core of our analysis lies in examining behavior of stable parts of phones, we also examined transformation on entire phones including both stable and transition frames. In-between the marked phone boundaries and centers, we align source and target frames proportionally in time. Table 4 summarizes the transformation results of the different approaches shown in Table 2 for learning and decoding with both stable and transition frames. Comparing the results in Tables 2 and 4, we see that all methods suffer in performance with the inclusion of transition frames. Moreover, the general decrease in average maximum mixture weight indicates that there are more heavy mixtures in the source and joint decoding, yielding higher error overall. Note that the phoneme classification is less effective on the transition frames, though still outperforms the classic acoustic GMM-based method with source classification. These results indicate that there is interest in treating stable parts and transitions of phones separately. For example, conversion can be carried out on stable parts and an interpolation method can be used on transitions.

Table 4. *Conversion Results for Phonetic & Acoustic GMM, Stable & Transition Frames.*

GMM (Q=28)	**MSE (dB)**	**Mean($w_{\max}$)**
P,Phoneme Separation	-4.87	--
P, X Decode	-4.40	0.62
A, X Decode	-4.43	0.67
P, XY Decode	-5.48	0.84
A, XY Decode	-5.80	0.89

5. Discussion & Future Work

Our work follows the common VC method modifying acoustic parameters of source speech in order to estimate target speech. The advantage in this type of approach to VC is that it allows for the estimated speech to keep fine structures in the original signal that are not necessarily captured in the conversion model. A perceptual evaluation of converted speech quality modifying the spectral envelope and pitch of source speech is carried out in [9]. It is noted that, while the synthesis model in this case can achieve a reasonable speech quality (using original target parameters, for example), the models for parameter conversion need improvement. In informal listening tests, we confirmed these observations and further found that, due to the degraded quality of the speech synthesized with converted parameters, it was difficult to make a clear distinction between the different decoding strategies presented in this paper.

The problems in learning and converting speaker-dependent parameters occur because time-alignment of individual source and target speech frames does not necessarily ensure alignment on an acoustic level. In future work, we will seek to use alternative methods to align speaker feature vectors. In particular, the results in this paper suggest that there is value in incorporating more context-dependent information in learning and conversion. Accordingly, our goal will be to use both temporal and contextual information to generate aligned speaker spaces with classes that are more acoustically homogeneous. This type of model generation relying on context-dependent information more closely follows methods for VC using adaptation of detailed speaker models [9]. We, however, will maintain an approach to VC using transformation directly on acoustic parameters of original source speech in order to achieve high quality speech synthesis.

6. Conclusion

We have shown in the context of GMM-based spectral conversion that one-to-many mappings exist and present a significant source of error in voice conversion. Furthermore, using context-dependent phonetic information in learning and classification for the transformation model alleviates the problem presented by one-to-many mappings, thus greatly improving the conversion results.

7. References

[1] Kain, A., and Macon, M., "Spectral voice conversion for text-to-speech synthesis," in Proc. of ICASSP '98, vol. 1, pp. 285-288.

[2] Sambur M.R., "Selection of Acoustic Features for Speaker Identification," IEEE Trans. on Acoust., Speech, Signal Processing, vol. 23., No. 2, April 1975.

[3] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H., "Voice Conversion through vector quantization," in Proc. of ICASSP '88, vol. 1, pp. 655-658.

[4] Narendranath, M., Murthy, H. A., Rajendran, S., and Yegnanarayana, B., "Transformation of Formants for Voice Conversion using Artifical Neural Networks," Speech Comm, vol. 16, issue 2, pp. 207-216, 1995.

[5] Stylianou, Y., Cappé, O., and Moulines, E., "Continuous probabilistic transform for voice conversion," IEEE Trans. On Speech & Audio Processing, vol. 6, Issue 2, pp. 131-142, 1998.

[6] Baudoin, G., and Stylianou, Y., "On the transformation of the speech spectrum for voice conversion," in Proc. of ICSLP '96, vol. 2, pp. 1405-1408.

[7] Mouchtaris, A., Agiomyrgiannakis, Y., and Stylianou, Y., "Conditional vector quantization for voice conversion," in Proc. of ICASSP '07, vol. 4, pp. 505-508.

[8] En-Najjary, T. "Conversion de voix pour la synthèse de la parole," Ph.D. diss., ENST Bretagne, France, April 2005.

[9] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai J., "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," IEEE Trans. On Audio, Speech and Language Processing, vol. 17., No. 1, pp. 66-83, 2009.