CONDITIONAL VECTOR QUANTIZATION FOR VOICE CONVERSION

A. Mouchtaris, Y. Agiomyrgiannakis, and Y. Stylianou

Multimedia Informatics Laboratory, Department of Computer Science, University of Crete and
Institute of Computer Science (FORTH-ICS)
711 10 Heraklion, Crete, Greece
{mouchtar, jagiom, yannis}@csd.uoc.gr

ABSTRACT

Voice conversion methods have the objective of transforming speech spoken by a particular source speaker, so that it sounds as if spoken by a different target speaker. The majority of voice conversion methods is based on transforming the short-time spectral envelope of the source speaker, based on derived correspondences between the source and target vectors using training speech data from both speakers. These correspondences are usually obtained by segmenting the spectral vectors of one or both speakers into clusters, using soft (GMM-based) or hard (VQ-based) clustering. Here, we propose that voice conversion performance can be improved by taking advantage of the fact that often the relationship between the source and target vectors is one-to-many. In order to illustrate this, we propose that a VQ approach namely constrained vector quantization (CVQ), can be used for voice conversion. Results indicate that indeed such a relationship between the source and target data exists and can be exploited by following a CVQ-based function for voice conversion.

Index Terms— Spectral analysis, speech synthesis, vector quantization, voice conversion.

1. INTRODUCTION

Voice conversion is a specific area of speech synthesis that focuses on synthesizing speaker identity. More specifically, in voice conversion the objective is to transform the voice of a particular (source) speaker, so that it sounds as the voice of a different (target) speaker. The most direct application of this area has been within the context of Text-To-Speech (TTS) synthesis, for adding new voices to a TTS system without the need of retraining the system for every new voice. More generally, voice conversion can find applications in other areas of speech processing where speaker identity is of importance, e.g. speech translation.

Current voice conversion methods are based on the assumption that the important features that characterize the speaker identity (and thus need to be transformed) are the spectral envelope (segmental feature) and pitch and time evolution with time (supra-segmental features). While pitch and time scaling on an average level can produce convincing transformation in the supra-segmental level, the transformation of the spectral envelope of one speaker into another (spectral conversion) is a more challenging problem. This is due to the fact that the spectral conversion algorithm must provide a useful result for every possible spectral envelope of the source speaker. In other words, the algorithm must be able to derive a generally applicable conversion function from the knowledge of some available speech data that are necessary in order to derive the proper conversion parameters.

A number of different approaches have been proposed for voice conversion. They all have in common some sort of segmentation of the source (and possibly the target) vector spaces, so that a meaningful relation can be derived between the source and target vectors. Early attempts were based on vector quantization (VQ) approaches, where a correspondence was derived between the source and target spectral envelope codebooks during the training procedure [1]. A problem with VQ methods has been the fact that the limited space of vectors available for synthesis introduces a degradation in the resulting speech quality. Gaussian Mixture Model (GMM) -based methods have been proposed [2, 3], which, due the continuous form of the GMM-based conversion function they propose, result in more natural synthesis. One issue with both VQ- and GMM-based conversion methods has been the need of training speech data that contain the same context, so that the source and target vectors can be correctly associated (parallel corpus). A generalization of GMM-based conversion methods that allows for a non-parallel speech training dataset has been proposed in [4].

In this paper we are interested to investigate the possibility of an one-to-many relationship between the source and target spectral vectors, and propose a training method that is able to exploit this property. The consequence of one-to-many relationships between the source speaker and the target speaker is that an optimal estimator will average the spectral envelopes of different sounds. Such an estimation will provide speech that is perceived as "muffled". The possibility of one-to-many relationship is important to be investigated, since in that case the mutual information between the source and target spaces could be exploited towards substantially improving the performance of current conversion methods. A VQ-type algorithm that can take advantage of this type of relationships is the Conditional Vector Quantization (CVQ) [5][6]. This method assumes an one-to-many relationship between the source and target vector spaces, and quantizes the vectors of the target space conditioned on the quantization of the source space. Here, we apply CVQ at the training dataset of a source and target speaker, and show that the best choice of the (limited) set of target vectors given a source vector results in significant improvement compared to conventional VQ-based conversion in terms of mean-squared error (MSE). Consequently, we do not propose here an improved solution to the problem of spectral conversion. This paper presents a first step towards showing that indeed one-to-many relationships exist between source and target vectors in the voice conversion context, and furthermore towards approaches that exploit such relationships for this problem.

This work has been funded by the Greek General Secretariat for Research and Technology, Program ΕΠΑΝ Code 05NON-EU-1, and by a Marie Curie International Reintegration Grant within the 6th European Community Framework Program.

2. VQ AND CVQ FOR VOICE CONVERSION

2.1. VQ-based Conversion

A popular method for spectral conversion based on VQ has been proposed in [1]. During the training phase of this algorithm, spectral vectors (e.g. LSFs, cepstral coefficients, etc.) from both the source and target speakers are available in an aligned form (e.g. from a parallel corpus). Initially, the source and target spaces are divided in clusters using VQ clustering, and each of the training source and target vectors is associated with one of the source and target clusters respectively. Then, using the fact that the source and target vectors are aligned in pairs, a histogram is created based on the number of occurrences of the target space clusters for each given source space cluster. This is in fact the probability of occurrence of each target space cluster given a source space cluster. During the conversion phase of the approach, each source vector is associated with one of the clusters of the source space, and the target vector is obtained by weighting all the target space centroids using the corresponding histogram that was derived during training for this particular source cluster. The method results in convincing conversion, however the fact that the converted space is limited (limited number of centroids weighted by a limited number of histograms) results in degraded audio quality. GMM-based conversion methods are based on a soft-clustering approach and use a continuous transformation function, thus improving the final audio quality in a large degree. However, here we focus our interest on VQ-based methods.

2.2. Conditional Vector Quantization

An estimator may not be able to benefit from the mutual information between the source space (X-space) and the target space (Y-space) when complicated one-to-many mappings take place. For example, a phoneme */a/* of the source speaker may be spoken in two different ways by the target speaker, each producing a sound with different spectral characteristics. In that case, an estimator will average these two spectra and provide a target spectrum that differs from both of them, producing speech of lower quality (usually "muffled"). The muffling effect could be avoided if either of these sounds is somehow chosen. Therefore, it is interesting to investigate and capture the one-to-many relationships that may be present in our training data.

The one-to-many relationships can be captured using Conditional Vector Quantization (CVQ). CVQ is based on two linked codebooks; a X-space codebook with M entries and a Y-space codebook with $M * K$ classes. Each entry of X-space codebook is linked to K entries of the Y-space codebook. The linking between the two codebooks is illustrated in Fig. 1.The design of the CVQ codebooks is made with a two-step procedure. First, the X-space vectors are quantized to M classes. Second, for each X-space class the Y-space vectors that correspond to this class are quantized to K classes [6]. Therefore, in practice, CVQ quantizes the residual of a VQ-based estimator inside each X-space class. If one-to-many relationships exist between the two spaces then the K CVQ codevectors that belong to the conditional $Y|X$-space will be allocated to capture these complex statistics. In other words, CVQ can provide insight to the Y-space classes which are mapped to a single X-space class (i.e. a phoneme). When one-to-one relationships are present, CVQ will simply encode the estimation residual.

2.3. CVQ-based conversion

Once CVQ has been applied for clustering the source and target vector spaces during the training phase, during the conversion phase each source vector is associated to a particular cluster of the source space. There exist some clusters in the target space that are associated with this particular source cluster. For example, if during training we assumed a 1-4 relationship in our data, this translates into 4 different centroids in the target space for each source cluster. In our initial approach, we suggest that the target vector during conversion could be one of those 4 centroids, although this approach will also suffer from the limited spectral variability problem encountered in VQ conversion. However, at this point it is not clear how to derive a voice conversion system based on CVQ that will be able to exploit the possibly present one-to-many relationships encountered during the training phase. In the remainder of the paper, we will choose the closest of these 4 centroids to the true target vector (not feasible in practice), in order to illustrate the possibilities of the CVQ approach in terms of mean-squared error when compared to the VQ-based conversion. The worst choice and the median distance for each of the 4 possible choices will also be given for better understanding the results.

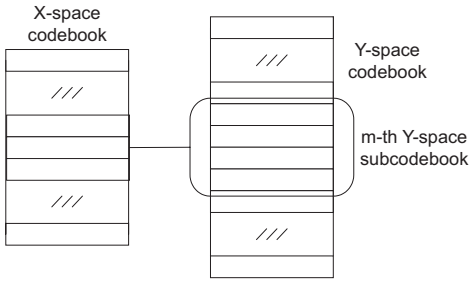

Fig. 1. CVQ codebooks.

3. RESULTS AND DISCUSSION

In this section, we are interested to test the best possible performance that can be achieved under the CVQ approach compared to the VQ-based conversion algorithm. In order to do this, we use a parallel training corpus of 2 speakers, containing 188 short sentences from each speaker (10 additional sentences were used for testing, i.e. deriving the results give later in this section). The sampling rate for the corpus is 16 kHz. The 188 training sentences were analyzed using 32 msec window with 75% overlapping, and the LSFs for each frame (22^{nd} order) were extracted. Each frame was classified as voiced or unvoiced using YIN [7], and only voiced frames were used for deriving the training vectors (since formant conversion is meaningful only for voiced speech frames). The procedure resulted in about 45,000 LSF vectors for each speaker. These were associated in pairs using a standard DTW algorithm, so that finally our training dataset consisted of 45,000 pairs of LSF vectors. For the VQ method, each of the source and target spaces was clustered separately, and then a histogram of occurrence for the target clusters for each source cluster was created. Various choices regarding the number of clusters was implemented as explained later in this section. The error measure used in this section is the mean-squared error (between LSF vectors) normalized by the initial distance between the reference and target speakers' LSFs, *i.e.*

$$\mathcal{E} = \frac{\frac{1}{N}\sum_{k=1}^{N}\left\|\vec{y}_k - \hat{\vec{y}}_k\right\|^2}{\frac{1}{N}\sum_{k=1}^{N}\left\|\vec{y}_k - \vec{x}_k\right\|^2},$$

where $\vec{x}_k$ is the source vector at instant k, $\vec{y}_k$ is the target vector at instant k, and $\hat{\vec{y}}_k$ denotes the estimated target vector using a spectral conversion method.

In Fig.2, we plot the Normalized MSE for the VQ, NLIVQ ([8],

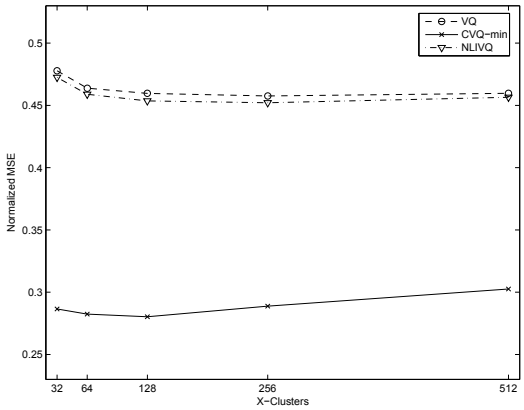

Fig. 2. Normalized MSE for various choices of clusters for the source vector space, for VQ, NLIVQ, and CVQ conversion.

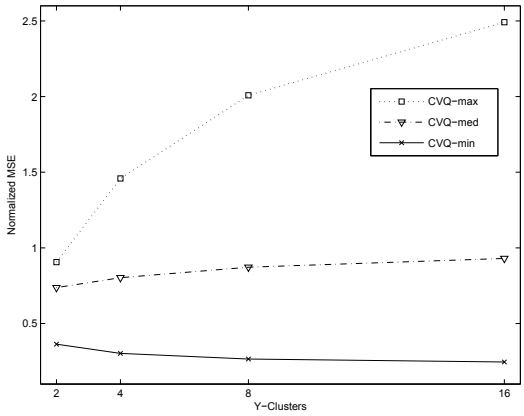

Fig. 3. Normalized MSE for various choices of clusters for the conditional target vector space, for CVQ conversion.

similar to the VQ-based conversion described earlier), and CVQ for various numbers of clusters of the source space (denoted as x-clusters). For CVQ, we obtain as the conversion result the conditional target centroid that is closest to the target vector (*i.e.* the best choice of target centroid, denoted in the figure as CVQ-min). In other words, we assign the source vector to one of the source space clusters, and for each of the corresponding target centroids we calculate its distance to the target vector and obtain the one with the minimal distance. For CVQ, the number of clusters of the target space is 4. Even with such a small number of available candidates, the error obtained with CVQ is about half compared to VQ. For 256 clusters of the source space, the MSE for CVQ is 0.2888 (best choice of target vector), while the MSE for VQ conversion is 0.4576. The corresponding MSE for the worst choice for CVQ conversion is 1.4277 (choosing the centroid that is farthest from the target vector), while the median MSE for the 4 possible choices (averaged over all testing vectors) is 0.8258. The fact that even with only 4 choices of target vectors we obtain such a large difference between the best and worst choice of vectors, is an indication that the variability is high in the clustered target space and indeed a one-to many relationship exists between the source and target spaces.

The observations made in the previous paragraph are further supported by the results in Fig. 3. In Fig. 3, we plot the Normalized MSE for the CVQ method, for 512 clusters in the source space, for various numbers of clusters in the target space (denoted as y-clusters in the figure). CVQ-min again indicates the case when the closest of the conditional target centroids is chosen for a particular source vector during testing (best choice as explained above), CVQ-max corresponds to the worst choice of centroid, and CVQ-med corresponds to the median distance of the possible choices (averaged over all testing vectors). This figure indicates that when increasing the number of clusters in the target space, the MSE increases greatly for the maximum distance case, while the best and median distances do not significantly vary. This is also an indication that there is great variability in the conditioned target space, which in turn implies a one-to-many relationship between the source and target vector spaces.

3.1. One-to-many relationship

Proving that the relationship between two spaces is one-to-many is not trivial. Large intra-class variations in the conditional target space could also be perceived as different classes. Many-to-many relationships can also manifest themselves as one-to-many relationships. Furthermore, the distortion measure affects the classification. For the task of speaker modification, we are interested to examine whether the spectral envelopes of the conditional target space have very different spectral characteristics (i.e. formants, spectral nulls). To do so, we conducted the following experiment: let $\vec{x}_n$,$\vec{y}_n$, $n = 1, ..., N$ be the N aligned source and target speaker vectors and $X_m = \{\vec{x}_n : Q(\vec{x}_n) = m\}$, $m = 1, ..., M$ be the set of X-space vectors quantized to the m-th X-space class, where $Q(\cdot)$ denotes the vector quantization procedure. Let $Y_m = \{\vec{y}_n : Q(\vec{x}_n) = m\}$ be the set of Y-space vectors that correspond to X_m. We quantized X_m and Y_m with 4 classes, and obtained the 4 representatives $\hat{\vec{x}}_{m,i}$, $\hat{\vec{y}}_{m,i}$, $i = 1, ..., 4$ respectively. Note that CVQ provides the same vectors $\hat{\vec{y}}_{m,i}$, $i = 1, ..., 4$ for each X-space class. These representative vectors provide insight to the structure of the X-space class and the corresponding conditional Y-space samples. Then we computed the maximum log-spectral distortion between all possible pairs of the vectors $\hat{\vec{x}}_{m,i}$ and $\hat{\vec{x}}_{m,i}$ to obtain the following *Diffusion Metrics*:

$$D_{x,m} = \arg\max_{i,j} \left\{ D(\hat{\vec{x}}_{m,i}, \hat{\vec{x}}_{m,j}) \right\},\ m = 1, ..., M \qquad (1)$$

$$D_{y,m} = \arg\max_{i,j} \left\{ D(\hat{\vec{y}}_{m,i}, \hat{\vec{y}}_{m,j}) \right\},\ m = 1, ..., M \qquad (2)$$

where $D(\cdot, \cdot)$ denotes the mean log-spectral distortion between two LSF vectors. The metric $D_{x,m}$ is a measure of the diffusion of the m-th X-space class, while $D_{y,m}$ is a metric of the diffusion of the Y-space samples that correspond to this class. The ratio between the two metrics provides the *Diffusion Number* of the corresponding class (intra-class diffusion):

$$R_m = \frac{D_{y,m}}{D_{x,m}},\ m = 1, ..., M. \qquad (3)$$

The Diffusion Number states the spread of the Y-space class relatively to the spread of the X-space class. If R_m is higher than one, then the variability that is encountered within the m-th conditional Y-space class is much higher than the variability encountered within the corresponding X-space class.

Fig. 4 shows the histogram of the diffusion metrics for M=512 X-space classes. Note that the diffusion metrics for the conditional Y-space are much higher than the diffusion metrics of X-space. Furthermore, in absolute terms, the $Y|X$-space diffusion metrics reveal a high variability on the conditional space: the mean distortion (diffusion metrics) is about 8 dB, and distortions above 10 dB or 12 dB are also common. Further insight is provided by Fig. 5 where the histogram of the Diffusion Number is shown. The variability within the conditional Y-space class is usually more than 3 times the variability

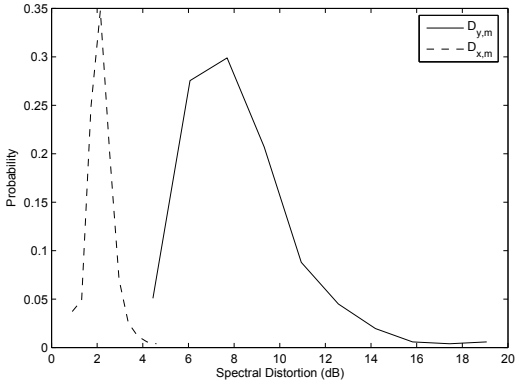

Fig. 4. Histograms of Diffusion Metrics

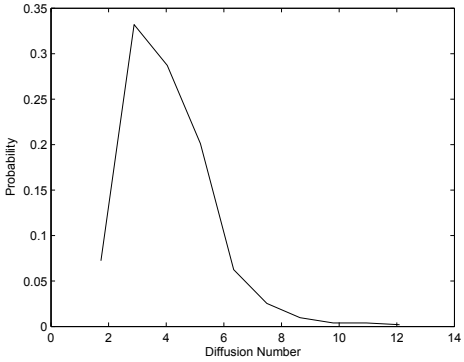

Fig. 5. Histogram of Diffusion Number

within each X-space class. For a considerable portion of X-space classes, the variability is more than 6 times larger. These observations indicate a one-to-many relationship when mapping X-space to Y-space, at least for some sound classes.

An example of the spectral envelopes $\hat{\vec{y}}_{m,i}$, $\hat{\vec{x}}_{m,i}$, $i = 1, ..., 4$ is given in Fig. 6. The upper four spectral envelopes (continuous lines) correspond to the conditional Y-space, while the lower four envelopes correspond to the X-space class. We observe that there is little variability in the X-space class while there is high variability of the conditional Y-space class. The X-space class contains a vowel with variations mainly at the first formant, while the $Y|X$-space class contains spectral envelopes that correspond to different sounds, since they differ on the number and location of the formants, as well as the spectral tilt. The example features diffusion metrics $D_{y,m}$ and $D_{x,m}$ equal to 11 dB and 1.66 dB, respectively. An average over this conditional space will produce a spectral envelope that doesn't resemble the characteristics of a clearly spoken vowel, resulting to the sound that is perceived as "muffled" which is the major source of degradation in current speaker conversion systems.

Knowledge regarding the one-to-many relationship can be used in order to produce a sequence of estimated target spectral envelopes that is perceived as a possible realization of the trajectories of the target speaker spectral envelopes. In this context, a possible sequence of spectral envelopes maybe recovered.

4. CONCLUSIONS

In this paper, we examined the problem of vector quantization (VQ) for voice conversion under the assumption that the relationship between the source and target spectral vector spaces is one-to-many. Conditional vector quantization (CVQ) was examined, which takes advantage of such relationships during the training procedure of a voice conversion method. Our initial results indicate that the assumption of such a relationship is a valid one, and that CVQ can indeed exploit this relationship and improve the conversion performance. In our future research we intend to focus on the issue of deriving a conversion algorithm based on these observations, in the direction of a continuous conversion function, so that speech quality issues that are inherent in VQ-based systems can be avoided.

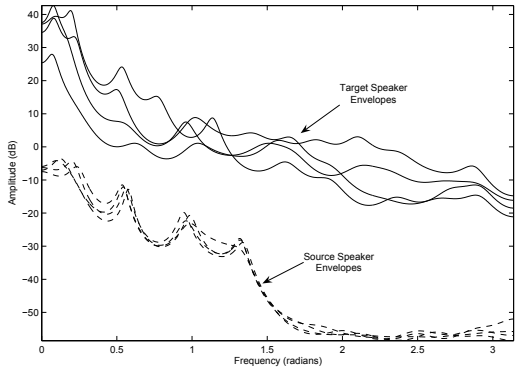

Fig. 6. An example of $\hat{\vec{x}}_{m,i}$ and $\hat{\vec{y}}_{m,i}$, $i = 1, 2, 3, 4$ for the m-th speech class.

5. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, New York, NY, April 1988, pp. 655–658.

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

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, May 1998, pp. 285–289.

[4] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 3, pp. 952–963, May 2006.

[5] Y. Agiomyrgiannakis and Y. Stylianou, "Coding with side information techniques for LSF reconstruction in voice over IP," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, 2005.

[6] Y. Agiomyrgiannakis and Y. Stylianou, "Conditional vector quantization for speech coding," *accepted to IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[7] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.

[8] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, 1992.