ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Modeling a Noisy-channel for Voice Conversion Using Articulatory Features

Bajibabu Bollepalli[1], *Alan W Black*[2], *Kishore Prahallad*[1]

[1]Speech and Vision Lab, International Institute of Information Technology, Hyderabad, India
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
bajibabu.b@research.iiit.ac.in, awb@cs.cmu.edu, kishore@iiit.ac.in

Abstract

In this paper, we propose modeling a noisy-channel for the task of voice conversion (VC). We have used the artificial neural networks (ANN) to capture speaker-specific characteristics of a target speaker which avoid the need for any training utterance from a source speaker. We use articulatory features (AFs) as a canonical form or speaker-independent representation of a speech signal. Our studies show that AFs contain a significant amount of speaker information in their trajectories. Suitable techniques are proposed to normalize the speaker-specific information in AF trajectories and the resultant AFs are used in voice conversion. The results of voice conversion evaluated using objective and subjective measures confirm that AFs can be used as a canonical form in nosiy-channel to capture speaker-specific characteristics of a target speaker.

Index Terms: voice conversion, articulatory features, noisy-channel model, speaker-independent representation.

1. Introduction

The problem of voice conversion is typically viewed as capturing an optimal mapping function between a source and a target speaker. To capture this mapping function, techniques rely on parallel data (i.e., source and target speakers record a set of same sentences) or non-parallel data with adaptation techniques [1, 2, 3, 4]. Saito *et. al.,* [5] have proposed a voice conversion method using a large amount of non-parallel data from a target speaker. However, this method still relies on a few parallel utterances for estimation of joint density model.

We prefer to view the problem of voice conversion as capturing speaker-specific characteristics and imposing these characteristics on an arbitrary source speech signal [6]. The problem of capturing speaker-specific characteristics can be viewed as modeling a noisy-channel [7]. Suppose C is a canonical form of a speech signal, a generic and speaker-independent representation of the message in speech signal passes through the speech production system of a target speaker to produce a surface form S. This surface form S carries the message as well as the identity of the speaker. One can interpret S as the output of a noisy-channel for the input C. Here, the noisy-channel is the speech production system of the target speaker.

The mathematical formulation of this noisy-channel model is –

$$\underbrace{argmax}_{S} p(S/C) = \underbrace{argmax}_{S} \frac{p(C/S)p(S)}{p(C)} \quad (1)$$

$$= \underbrace{argmax}_{S} p(C/S)p(S) \quad (2)$$

as $p(C)$ is constant for all S. Here $p(C/S)$ could be interpreted as production model. $p(S)$ is the prior probability of S and it could be interpreted as the continuity constraints imposed on the production of S. It could be seen analogous to a language model of S.

Figure 1: Capturing speaker-specific characteristics as a speaker-coloring function.

In this work, $p(S/C)$ is directly modeled as a mapping function between C and S using artificial neural networks (ANN). The process of capturing speaker-specific characteristics and its application to voice conversion is explained below:

We derive two different representations C and S from a speech signal with the following properties: Let, C be a canonical form of speech signal, i.e., a generic and speaker-independent form - approximately represented by articulatory features (AFs) extracted from speech signal. Let S be a surface form represented by Mel-cepstral coefficients (MCEPs). If there exists a function $\Omega(.)$ such that $S' = \Omega(C)$, where S' is an approximation of S- then $\Omega(C)$ can be considered as specific to a speaker. The function $\Omega(.)$ could be interpreted as speaker-coloring function. We treat the mapping function $\Omega(.)$ as capturing speaker-specific characteristics. It is this property of $\Omega(.)$, we exploit for the task of voice conversion. Fig. 1 depicts the concept of capturing speaker-specific characteristics as a speaker-coloring function.

2. Encoder: Extraction of articulatory features

Following Black *et. al.,* [8], the articulatory features (AFs) used in this work represent the characteristics of speech production process, which include manner of articulation, place of articulation, height of vowel, etc. as shown in Table 1. We have used eight different articulatory properties, as tabulated in the first column of Table 1. Each articulatory property has a different number of classes, where each class is denoted by a separate dimension in AF space. For example, vowel length has four classes – short, long, schwa and diphthong. To represent these four classes, we have used four bits. The dimension of an AF vector is 26, which is equal to the total number of bits present in the third column of Table 1.

Pattern recognition techniques like artificial neural networks, and support vector machine classifiers (SVMs) are typically used for the estimation of AFs from acoustic signal [9]. These methods build a separate articulatory classifier for each

Table 1: Eight articulatory properties, each property has different classes and the number of bits required to represent each property.

Articulatory properties	classes	# bits
Voicing	+voiced, -voice	1
Vowel length	short, long, diphthong, schwa	4
Vowel height	high, mid, low	3
Vowel frontness	front, mid, back	3
Lip rounding	+round, -round	1
Consonant type (Manner)	stop, fricative, affricative, nasal, liquid, approximant	6
Place of articulation	labial, velar, alveolar, palatal, labio-dental, dental, glottal	7
Silence	+silence, -silence	1

AF type. Models are trained to predict the presence or absence of an AF type, and finally the outputs of these classifiers are concatenated to form an AF vector.

In this work, we rely on building an ANN mapper which maps an MCEP vector to an AF vector. Such mapper uses lesser number of parameters and also preserves the dependencies or correlations among AFs. The structure of the ANN model used is $25L50N20L50N26L$, where the integer value indicates the number of nodes in each layer and L / N indicates the linear or nonlinear activation function.

3. Are articulatory features speaker specific?

In recent years, articulatory features have been used for automatic speech recognition (ASR) with the aim of better pronunciation modeling [9], better co-articulation modeling, robustness to cross speaker variation and noises, multi-lingual [10] and cross-lingual portability of systems, language identification [11] and expressive speech synthesis [8]. In these studies, often the articulatory features derived from the acoustics are treated as generic or speaker-independent representations of the speech signal. We wanted to investigate how much speaker-specific information is left out in the AFs. To answer this, we conducted a speaker recognition experiment using AFs.

In this experiment, we built a speaker identification system (SID) on 630 speakers of TIMIT database using AFs and MCEPs. To extract spectral features from the speech signal, an excitation filter model of speech was applied, and MCEPs were extracted using a frame size of 25 ms with a fixed frame advance of 5 ms. AFs were extracted from MCEPs by building an encoder for each speaker as explained in Section 2. The performance of the AF based SID system was compared with that of the MCEP based SID system. We hope to obtain an identification performance above chance level for the set of 630 speakers, if the AFs capture the identity of a speaker.

A Gaussian mixture model (GMM) was used to model the distribution of features of a given speaker. Each speaker was modeled by 32 mixtures and the models were trained using expectation maximization (EM) algorithm, with an initial model trained using the k-means algorithm.

Table 2: Accuracy of the (%) speaker identification system using MCEPs and AFs

Features	ACC (%)
MCEPs	100
AFs	85.24

3.1. Results

All the SX and SI wave-files in each speaker's directory of TIMIT database were concatenated to form a single utterance (of approximately 25 seconds duration). Feature vectors (AFs/MCEPs) were extracted on this utterance to build a speaker model. Two utterances in SA directory of each speaker were concatenated to form a test utterance. To compute the accuracy, each test utterance was matched against all of the 630 speakers. The accuracy (ACC) of the speaker identification system was defined as the percentage of identifications which are correct.

Table 2 shows the performance of the SID system using MCEPs and AFs. The accuracy of the SID system using MCEPs is 100%. The accuracy of the SID system using AFs is about 85%, which is far above the chance level. This indicates that AFs do capture sufficient amount of speaker information in their covariance matrices, but may not be as good as that of MCEPs.

This raises the question – how one could normalize the speaker information in AFs, so that AFs act as a speaker-independent representation of the speech signal. Such representation can be used as a canonical form in the noisy-channel model for capturing speaker specific characteristics.

3.2. Normalization of speaker specific information

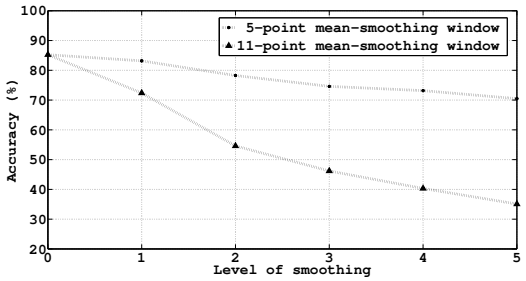

Figure 2: Speaker identification accuracies for different levels of smoothing by a 5-point and an 11-point mean-smoothing window. Level 'k' corresponds to applying mean-smoothing window 'k' times.

In order to normalize the speaker specific information in AF streams, we performed mean smoothing of the AF trajectories with a 5-point and an 11-point window. The idea was to smooth the correlations among the samples in the AF trajectories to normalize the effect of speaker-specific characteristics.

Fig. 2 shows the speaker identification performance after applying the mean-smoothing iteratively for five times. It is observed that the performance of SID system decreases with every iteration of mean-smoothing and more so for the 11-point window spanning 225 milliseconds.

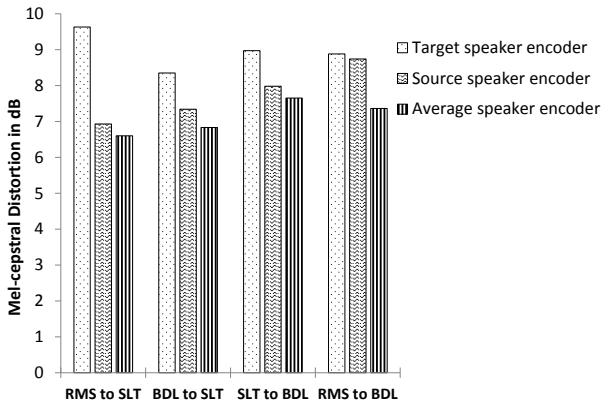

Figure 3: Plot of MCD scores obtained between different speaker pairs.

4. Use of smoothed articulatory features for voice conversion

4.1. Training target speaker's model

Given the utterance from a target speaker T, the corresponding canonical form C_T of the speaker was represented by AFs. To alleviate the effect of speaker characteristics, the AFs undergo a normalization technique such as smoothing, as explained in Section 3.2. The surface form S_T was represented by traditional MCEP features, as it would allow us to synthesize using the MLSA synthesis technique. The MLSA synthesis technique generates a speech waveform from the transformed MCEPs and $F0$ values using pulse excitation or random noise excitation [12]. An ANN model was trained to map C_T to S_T using the backpropagation learning algorithm by minimizing the Euclidean error $||S_T - S'_T||$, where $S'_T = \Omega(C_T)$.

4.2. Conversion process

Once the target speaker's model is trained, it can be used to convert C_R to S'_T where C_R is the canonical form from an arbitrary source speaker R. To get the canonical form for any arbitrary source speaker we could follow any of the three methods below: The process to build any of the encoders below is explained in Section 2.

1. Use source speaker encoder. This requires building an encoder specific to a source speaker, and hence a large amount of speech data (along with transcription) from source speaker is required.

2. Use target speaker encoder. This maps MCEPs of an arbitrary source speaker onto AFs using target speaker's encoder.

3. Use average speaker encoder. This maps MCEPs of an arbitrary source speaker onto AFs using an average speaker encoder which is trained using all speakers' data except that of source and target speakers. Since an average model is used to generate AFs, a form of speaker normalization takes place on AFs even before smoothing is applied.

4.3. Validation

By using the above three methods, we predicted the AFs for three source speakers SLT, BDL and RMS. The AFs were smoothed to normalize speaker-specific information. Smoothed AFs were mapped onto the BDL and SLT speaker-specific models. To test the effectiveness of the VC model, we computed the

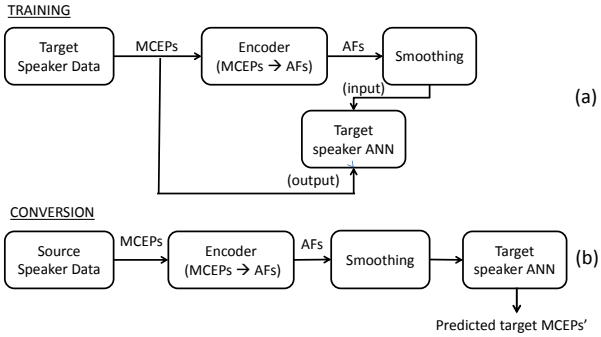

Figure 4: Flow-chart of training and conversion modules of a VC system capturing speaker-specific characteristics.

Mel-cepstral distortion (MCD) between predicted MCEPs and actual MCEPs. MCD is a standard measure used in speech synthesis and voice conversion evaluations [1]. Fig. 3 shows the MCD scores obtained using the above three methods. We observe that average speaker encoder gives a lesser MCD score compared to other methods. This justifies that the use of an average speaker encoder generates normalized AFs, and smoothing the AF trajectories further help in realizing the speaker-independent form. Rest of the experiments were carried out using average speaker encoder to get canonical form for any arbitrary source speaker.

4.4. Experiments on multiple speaker database

To test the validity of the proposed method, we conducted experiments on other speakers' database from the CMU ARCTIC set, such as RMS, CLB, AWB, and KSP. Fig. 4.(a) shows the block diagram for the training process and Fig. 4.(b) shows the block diagram for conversion processing. Table 3 provides the results for mapping C_R (where R = BDL, RMS, CLB, AWB, KSP, SLT voices) onto the acoustic space of SLT and BDL.

Table 3: *MCD scores obtained between multiple speaker pairs with SLT and BDL as target speakers. Scores in parenthesis are obtained using parallel data.*

Source speakers	Target speakers	
	SLT	BDL
SLT	-	7.563 (6.709)
RMS	6.604 (5.717)	7.364 (6.394)
AWB	6.797 (6.261)	7.731 (6.950)
KSP	7.808 (6.755)	8.695 (7.374)
BDL	6.637 (5.423)	-
CLB	6.339 (5.380)	7.249 (6.172)

In Table 3 the performance of voice conversion models built following a noisy-model approach is compared with that of traditional model using parallel data. Here, we implemented the voice conversion system using parallel data using artificial neural networks as explained in [6]. MCD scores indicate that use of parallel data performs better than the noisy-channel model approach. The use of parallel data allows to capturing an explicit mapping function between a source and a target speaker. The approach of the noisy-channel model captures target speaker-specific characteristics which could be later imposed on any source speaker. This approach provides an MCD in the range of 6.3 to 8.6. The focus in this paper is to ob-

Table 4: Subjective evaluation of voice conversion models built by using parallel data and Noisy-channel model.

Transformation using	SLT to BDL	BDL to SLT
Parallel data	3.34	3.58
Noisy-channel model	3.14	3.40

tain a better transformation of spectral features. Hence, we use the traditional approach of $F0$ transformation as used in GMM based voice transformation [1].

4.4.1. Subjective evaluation

We have also performed perceptual tests whose results are provided in Table 4 for mean opinion scores (MOS) in the scale of 1 to 5 (5:Excellent, 4:Good, 3:Fair, 2:Poor, 1:Bad). For the listening tests, we chose 10 utterances randomly from the two transformed pairs (SLT to BDL and BDL to SLT). Fifteen listeners participated in the evaluation tests. The MOS scores in Table 4 are averaged over fifteen listeners. By observing the MOS scores, one could say that the noisy-model approach does capture speaker-specific characteristics of the target speaker. The transformed waveforms are available at *http://researchweb.iiit.ac.in/∼bajibabu.b/vc_evaluation.html*.

By using the smoothed AFs, we can transform any arbitrary speaker onto a predefined target speaker without the need of any utterance from a source speaker in training the voice conversion model. This indicates that the methodology of training an ANN model to capture speaker-specific characteristics for voice conversion could be generalized over different datasets.

4.5. Cross-lingual voice conversion

Cross-lingual voice conversion is a task where the language of the source and the target speakers is different. We employ the ANN model which captures speaker-specific characteristics for the task of cross-lingual voice conversion. We performed an experiment to transform the voice of a Tamil and a Telugu speaker into a male voice of an English speaker (US male - BDL). Our goal here is to transform two speaker voices to BDL voice and hence the output will be as if BDL were speaking in Tamil and Telugu, respectively. We make use of BDL models built in Section 4 to capture speaker-specific characteristics. Five utterances from two speakers were transformed into BDL voice. We then performed the MOS test and the similarity test to evaluate the performance of this transformation. Table 5 provides the MOS and similarity test results averaged over all listeners. There were five native listeners of Telugu, and Tamil who participated in the evaluation tests. The MOS scores in Table 5 indicate that the intelligibility of the transformed voice was not high. The similarity tests indicate how close the transformed speech bear the target speaker characteristics. These tests indicate that cross-lingual transformation could be achieved using ANN models, and the output possesses the characteristics of BDL voice.

5. Summary and conclusions

In this paper, we have shown that it is possible to build a voice conversion by capturing speaker-specific characteristics of a speaker (noisy-channel model). We have used an ANN model to capture the speaker-specific characteristics. Such a model does not require any speech data from source speakers and hence could be considered as independent of source speaker. We have used AFs to represent the canonical form of a speech signal. Using speaker identification experiments, we have shown that AFs contain a significant amount of speaker specific information. It is also shown that speaker information in AFs could be normalized by smoothing iteratively. Our results indicate that AFs can be used as a canonical form of the speech signal in the noisy-channel model to capture speaker-specific characteristics for voice conversion. An effective process of normalization or transformation of AFs for cross-lingual voice conversion have to be investigated further.

Table 5: Subjective evaluation of cross-lingual voice conversion models.

Source Speaker (Lang.)	Target Speaker (Lang.)	MOS	Similarity test
Speaker1 (Telugu)	BDL (Telugu)	1.85	2.40
Speaker2 (Tamil)	BDL (Tamil)	2.00	2.50

6. References

[1] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Lang. Process, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.

[2] A. R. Toth and A. W. Black, "Using articulatory position data in voice transformation," in Proc. 6th ISCA Workshop Speech Synth. (SSW6), Bonn, Germany, Aug. 2007, pp. 182-187.

[3] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and S. Narayanan, "Text-independent voice conversion based on unit selection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Toulouse, France, May 2006, vol. 1, pp. 81-84.

[4] A. Mouchtaris, J. V. Spiegal, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. Audio, Speech Lang. Process., vol. 14, no. 3, pp. 952-963, May 2006.

[5] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Probabilistic unification of joint density model and speaker model for voice conversion," in Proc. of Interspeech, 2010, pp. 1728-1731.

[6] S. Desai, A. W. Black, B. Yegnanarayana, K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," IEEE Trans. Audio, Speech Lang. Process., vol. 18, no. 5, pp. 954-964, Jul. 2010.

[7] K. Prahallad, "Automatic building of synthetic voices from audio books," PhD dissertation, CMU, Pittsburgh, Jul. 2010.

[8] A. W. Black et al., "Articulatory features for expressive speech synthesis", Proc. of IEEE Int. Conf. Acoust., Speech, Signal Process, Kyoto, Japan, 2012.

[9] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition:2006 JHU summer workshop final report," http://www.clsp.jhu.edu/ws2006/groups/afsr/documents /WS06AFSR_final_report.pdf , 2008.

[10] S. St ker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2003, vol. 1, pp. 144-147.

[11] Abhijeet Sangwan, Mahnoosh Mehrabani, John H. L. Hansen, "Language Identification using a Combined Articulatory Prosody Framework," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2011, pp. 4400-4403.

[12] S. Imai, "Cepstral analysis synthesis on the Mel frequency scale," in Proc. of IEEE Int. Conf. Acoust., Speech, Signal Process., Boston, MA, Apr. 1983, pp. 93-96.