PHONEME INDEPENDENT HMM VOICE CONVERSION

Winston Percybrooks[⋆†] *Elliot Moore*[⋆] *Correy McMillan*[⋆]

[⋆] Georgia Institute of Technology
Electrical and Computer Engineering
210 Technolgy Circle, Savannah, GA 31407
[†] Universidad del Norte
Electrical and Electronics Engineering
Barranquilla, Colombia

ABSTRACT

This paper presents a voice conversion algorithm based on Hidden Markov Models that does not requires explicit phonetic labeling of the input speech. Additionally, the proposed voice conversion algorithm also uses an excitation estimation algorithm previously presented by the authors to achieve higher speech quality without compromising speaker identity conversion. The performance of the proposed algorithm was compared, using listening tests, with the performance of a recent voice conversion algorithm based on HMM but requiring phonetic labeling. The proposed algorithm was found to achieve equivalent identity conversion scores while improving the perceived quality of the converted speech. Thus, the proposed algorithm was found as a viable alternative for conversion applications where phonetic labeling is not practical.

Index Terms— Phoneme independent, HMM, voice conversion, ABX, MOS

1. INTRODUCTION

Speech is a basic communication mean for human beings. Besides the explicit information, encoded in a particular language, that is transmitted everytime a person speaks, there is a plethora of non-linguistic cues that can be interpreted by the listeners. In particular, humans are able to associate a voice with the corresponding speaker. Consequently, there is an interest on developing methods to manipulate the perceived identity of a given voice without altering any other information that it may contain. Several practical applications have been proposed, among them: personalization of text-to-speech systems [1, 2], movie dubbing [3], foreign language learning [4], and as a component in speech-to-speech translation systems [5].

Voice conversion (VC) systems aim to transform segments of speech from a given source speaker so that it can be identified as spoken by a different target speaker. This conversion involves transforming spectral features as well as prosodic features like pitch and speaking rate [6]. Well stablished VC algorithms focus on transforming short-time vocal tract spectral features using statistical methods based on Gaussian Mixture Models (GMM) [1, 7]. Furthermore, results in [8, 9] showed that the source signal (e.g. linear prediction residual) in the classic source-filter model of speech production also contains valuable information, not only for speaker identification but also for speech naturalness. Recent work has attempted to improve on [8, 9] by borrowing ideas from speech synthesis using Hidden Markov Models (HMM) to map source to target vocal tract features.

Of particular interest for the work presented here are the ideas shown on [10], where a HMM-based system is proposed as an incremental modification of the one presented in [11]. While most VC systems in the literature deal with conversion between a single source and target speakers at a time, the main contribution of [10] is the creation of an speaker-independent voice conversion framework, where the source speaker can be changed with little or no extra training, broadening the range of possible applications for VC. However, because it requires phonetic labeling of the input speech, the system in [10] is still not easily applicable in scenarios of cross-language conversion with no bi-lingual training data, where a phonetic match between source and target language is not practical or possible. The system in [10] will be referred here as Phoneme Dependent - HMM (PD-HMM).

The main contribution of the work presented here is then to propose an HMM-based VC algorithm that requires no explicit phonetic labeling, making it more suitable for cross-language applications, while achieving at least the same level of performance in identity conversion and speech quality than the PD-HMM system. This system will be called Phoneme Independent - HMM (PI-HMM). In addition to propose an alternative way to use HMM to perform voice conversion,

Note: This material is based upon work supported by the National Science Foundation under Grant No. 1116475. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

the PI-HMM also uses an algorithm for excitation estimation proposed by the authors in [12], and which has been determined to improve both identity conversion and speech quality on GMM-based conversion systems [12, 13].

2. PHONEME-DEPENDENT HMM VOICE CONVERSION (PD-HMM)

Figure 1 shows a diagram of the Phoneme-dependent HMM VC system presented in [10]. For the training section, speech from multiple speakers is used to train speaker-independent phoneme HMMs for the particular language involved in the conversion, in the same way that would be done for a speaker-independent automatic phoneme recognizer. Training data from the target speaker is then used to adapt the speaker-independent HMM models into target speaker models using Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR) [14] and Maximum A Posteriori (MAP) estimation [15]. Both sets of models, speaker-independent and target-adapted, are stored for use in conversion. For conversion, first Mel-cepstrum, phoneme and F0 sequences are extracted from the source speaker speech. The phoneme sequence is obtained using the speaker-independent models and the Mel-cesptrum coefficients and then used for computing the F0 sequence using adaptive F0 quantization. The converted speech features (Mel-cepstrum coefficients) are generated from the target-adapted HMM model using the phoneme and F0 sequences from the previous step and Maximum Likelihood (ML) criterion. Finally, the converted speech is synthesized using a Mel-Log Spectrum Approximation (MLSA) filter.

3. PHONEME-INDEPENDENT HMM VOICE CONVERSION (PI-HMM)

The proposed Phoneme-Independent HMM conversion algorithm uses Linear Prediction (LP) filtering for analysis and synthesis, and is composed by two main stages. First, the spectral features, i.e. Linear Spectral Frequencies (LSF), from the source speaker are converted into the target speaker feature space using a neutral HMM and speaker adaptation. Second, the corresponding target excitation signal for synthesis is estimated from the converted LSF vectors using the algorithm presented in [12]. The next subsections give more details about these stages.

3.1. Conversion of vocal tract features

The first stage on the PI-HMM model for both training and coversion modes is a pitch-synchronous LP analysis, generating a sequence of LSF and corresponding inverse filtering excitations from the input speech. Using as many training speakers as available, a speaker-independent HMM is created using the LSF sequences as observations and GMM as each state distribution. Contrary to the PD-HMM algorithm from Section 2, on this single HMM states do not correspond to phonemes, but rather to acoustical units determined dynamically and automatically from the training data. Several HMM with different sizes, state distribution and inter-state connection limitation (fully-connected, left-to-right, etc) are trained and Maximum Likelihood (ML) criterion is used to select the model that better matches the training data. The resulting model is called 'neutral model' because it is specific to a particular language but not to a particular speaker. The neutral model is used next to classify the training LSF sequences into the underlaying states and create an LSF Vector Quantizing (VQ) table for each state. The neutral HMM model and VQ table are then stored for use in conversion mode. Additionally, as is the case for the PD-HMM algorithm, a target-dependent HMM is created using CSMAPLR to adapt the HMM parameters. The state-associated VQ tables are then updated for the target-specific model by classifying only the target speaker training data with the new model. Figure 2 illustrates the training process.

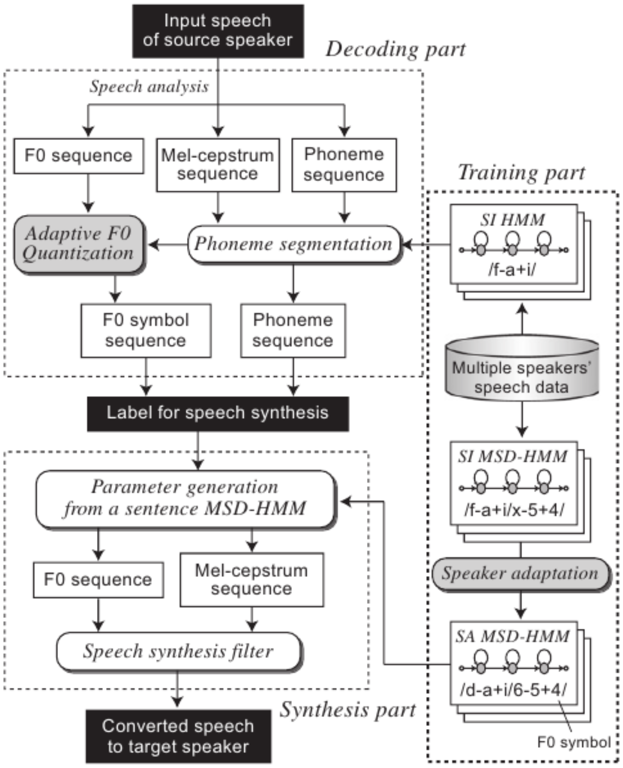

Fig. 1. Diagram for PD-HMM VC [10].

During conversion mode, the sequence of source speaker LSF computed from the analysis stage is then used in conjuction with the neutral model to generate the sequence of states with maximum likelihood (ML). That sequence of states is then input to the target-specific HMM to compute a converted

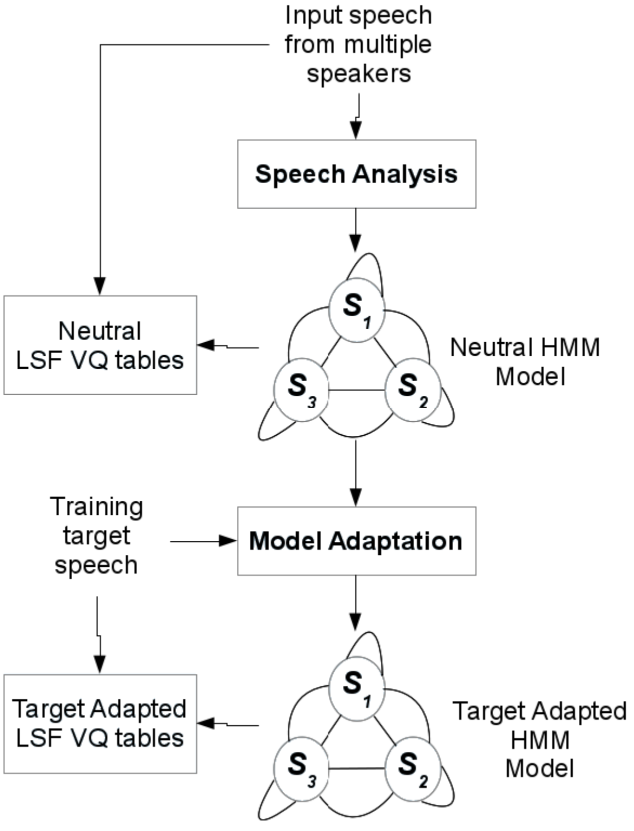

Fig. 2. Diagram for PI-HMM VC training.

sequence of target LSF vectors by applaying again the ML criterion. The resulting sequence of converted LSF vectors is the used as input for both, LP synthesis and the excitation estimation algorithm described on the next section.

3.2. Excitation estimation

During conversion, the converted spectral features (LSF) obtained during the previous stage are then used to estimate the corresponding excitation signal for the LP synthesis filter. This process involves a separate HMM model trained exclusively on target speaker data using LSF sequence as observations and the corresponding quantized inverse LP filtering excitations as hidden states. The training and estimation processes were already detailed by the authors on [12], an illustration can be seen on Figure 3.

4. EXPERIMENTAL SETUP

Subjective testing was used to compare the performance of the two VC systems under consideration. The Entropic Latino-40 speech database was used as source of data. The Latino-40 database contains speech from 40 speakers, 20 males and 20 females, of Latin American Spanish. For each speaker 125 sentences were chosen randomly from a pool of 13,000 sentences from Latin American newspapers, so each speaker has a different set of data in the database. For the tests presented here, 4 speakers, 2 males and 2 females, were chosen and for each speaker the pool of 125 available sentences was randomly divided into 100 sentences for training and 25 for testing. A series of listening tests were used to subjectively evaluate two main performance metrics: identity conversion and synthesis quality. The same group of 20, college age, native Latin American Spanish speakers participated as listeners on each listening test.

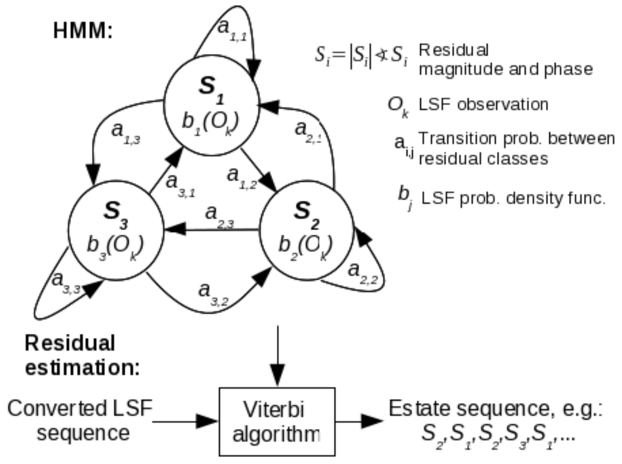

Fig. 3. Excitation estimation algorithm for PI-HMM VC.

An ABX test was done to evaluate how well each system performed identity conversion. For this ABX test, listeners had to rate 10 examples of conversions for each conversion case (i.e., Male-to-Male (M-M), Female-to-Female (F-F), Male-to-Female (M-F), Female-to-Male (F-M)) and VC system under test (i.e., PD-HMM or PI-HMM), according to the ABX scale (1, no conversion, to 5, perfect conversion into target).

A MOS test was carried out to evaluate converted speech quality from both systems, using the standard MOS scale that goes from 1 = Very poor quality to 5 = Excellent quality. For this MOS test, listeners were presented 10 sentences for each conversion case and each VC system. Additionally, the listeners also rated 10 original, unconverted target speaker sentences in order to give a reference value for the perceived quality of the original database recordings.

5. TESTS AND RESULTS

Table 1 summarizes the average ABX results for both PD-HMM and PI-HMM systems. Statistical analysis of the results showed that there is no significant difference, at the 1% or 5% level, for the two systems. This initial results indicates that the combination of a Phonetic-Independent Speaker Adapted HMM conversion for vocal tract features plus HMM estimation for the excitation signal is able to compensate for

Table 1. Average ABX results for every type of conversion. Male-to-Male (M-M), Female-to-Female (F-F), Male-to-Female (M-F), Female-to-Male (F-M).

Test type	**Avg. score PD-HMM**	**Avg. score PI-HMM**	**Difference**
M-M	3.98	4.00	0.02
F-F	3.95	4.02	0.07
M-F	4.01	3.97	-0.04
F-M	4.12	4.10	-0.02
Overall	4.01	4.02	0.01

Table 2. Average MOS results for every type of conversion. Male-to-Male (M-M), Female-to-Female (F-F), Male-to-Female (M-F), Female-to-Male (F-M).

Test type	**Avg. score PD-HMM**	**Avg. score PI-HMM**	**Difference**
M-M	3.77	3.88	0.11
F-F	3.72	3.85	0.13
M-F	3.62	3.91	0.29
F-M	3.65	3.83	0.18
Overall	3.69	3.87	0.18

the absence of phonetic labeling (and the knowledge about a particular language phonetics that it requires) on the input data as required by the PD-HMM algorithm. PI-HMM thus gives comparable performance on speaker identity evaluation while imposing less requirements on the training and testing speech data.

Regarding speech quality evaluation, Table 2 summarizes the average MOS results for each system and each type of conversion. As a reference, the average MOS score for the original recordings without conversion was found to be 4.92, confirming that the original database data is of high quality recordings. Taking into account that a difference of 1 point or more on the MOS scale is considered a very significant difference in perceived quality, the fact that both VC systems achieved MOS scores lower that 4.0 on every case while the original recordings were evaluated with a 4.92, indicates that the VC processing is still causing significant degradation compared with the input speech. Nonetheless, statistical analysis shows that the PI-HMM system achieves significant quality improvement at the 1% level for cross-gender conversion cases and 5% level for same-gender conversions with respect to the PD-HMM system. Arguably this could attributed to the excitation estimation stage adding some extra spectral details to the converted speech, which results on a higher perceptual naturalness and quality.

As an additional note, altought care must be taken because of the different experimental setups, the ABX and MOS results presented here when compared with the results in [12] and [10] seem to confirm the advantages of HMM-based VC systems versus traditional GMM-based systems. In particular, HMM-based systems are confirmed to achieve more consistent quality scores between same-gender and cross-gender conversions as opposed to GMM-based systems where there is a significant dip in quality for cross-gender conversions [12, 10].

6. CONCLUSIONS

The work presented here compares two HMM-based algorithms for voice conversion. The first algorithm, PD-HMM, requires a phonetic labeling step of the input speech, which implies an explicit knowledge of the phonetics of the language involved in the conversion. By droping that phonetic requierement, the proposed PD-HMM algorithm aims to be more suitable to application scenarios where it may be not practical to use explicit language phonetics, for example when source and target speakers are using different languages and thus a different set of standard phonemes. The results from subjective tests indicate that the proposed PI-HMM system achieves equivalent identity conversion performance than the PD-HMM according to ABX scores. On the speech quality metric, the PI-HMM algorithm achieves results statistically superior to the PD-HMM algorithm on the MOS scale, althought still showing considerable quality degradation with respect to the original unconverted recordings. According to the results presented here, future work will be focused on two aspects: improving quality scores to bring them closer to the quality of the input speech and apply the PI-HMM algorithm to cross-language conversion scenarios.

7. REFERENCES

[1] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 1998, pp. 285–289.

[2] W. Zhang, L.Q. Shen, and D. Tang, "Voice conversion based on acoustic feature transformation," in *6th National Conference on Man-Machine Speech Communications*, 2001.

[3] O. Turk and L. Arslan, "Subband based voice conversion," in *Internat. Conf. Spoken Lang. Process.*, 2002, pp. 289–292.

[4] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on gmm and straight," in *European Conf. Speech Comm. and Tech.*, 2001.

[5] H. Hoge, "Project proposal tc-star - make speech to speech translation real," in *Internat. Conf. Lang. Resources and Evaluation*, 2002.

[6] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Comm.*, vol. 16, pp. 165–173, 1995.

[7] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Audio, Speech and Language Process.*, vol. 6, pp. 131–142, 1998.

[8] A. Kain, *High resolution voice transformation*, Ph.D. thesis, OGI school of science and engineering, 2001.

[9] A. Kain and M.W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 2001, pp. 813–816.

[10] T. Nose and T. Kobayashi, "Speaker-independent hmm-based voice conversion using adaptive quantization of the fundamental frequency," *Speech Comm.*, vol. 53, pp. 973–985, 2011.

[11] T. Nose, Y. Ota, and T. Kobayashi, "Hmm-based voice conversion using quantized f0 context.," *IEICE Trans. Inf. Systems*, vol. I, pp. 2483–2490, 2010.

[12] W. Percybrooks and E. Moore, "A hmm approach to residual estimation for high resolution voice conversion," in *INTERSPEECH*, 2012.

[13] W. Percybrooks and E. Moore, "Voice conversion with linear prediction residual estimation," in *IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 2008.

[14] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *INTERSPEECH*, 2006.

[15] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Audio, Speech and Language Process.*, vol. 2, pp. 291–299, 1994.