DESIGN AND EVALUATION OF A VOICE CONVERSION ALGORITHM BASED ON SPECTRAL ENVELOPE MAPPING AND RESIDUAL PREDICTION

Alexander Kain and Michael W. Macon

Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute
20000 NW Walker Road, Beaverton, OR 97006, USA
kain@cse.ogi.edu, macon@ece.ogi.edu

ABSTRACT

The purpose of a voice conversion (VC) system is to change the perceived speaker identity of a speech signal. In this paper, we propose a new algorithm based on converting the LPC spectrum and predicting the residual as a function of the target envelope parameters. We conduct listening tests based on speaker discrimination of same/difference pairs to measure the accuracy by which the converted voices match the desired target voices. To establish the level of human performance as a baseline, we first measure the ability of listeners to discriminate between original speech utterances under three conditions: normal, fundamental frequency and duration normalized, and LPC coded. Additionally, the spectral parameter conversion function is tested in isolation by listening to source, target, and converted speakers as LPC coded speech. The results show that the speaker identity of speech whose LPC spectrum has been converted can be recognized as the target speaker with the same level of performance as discriminating between LPC coded speech. However, the level of discrimination of converted utterances produced by the full VC system is significantly below that of speaker discrimination of natural speech.

1. INTRODUCTION

The goal of voice conversion (VC) is to modify a *source* speaker's utterance to sound as if a *target* speaker had spoken it. Its uses include customization of text-to-speech systems (e.g. , to speak with a desired voice or to read out email in the sender's voice), as well as entertainment and security applications. To measure the performance of a VC system, the output must be evaluated by listening tests, especially when considering naturalness and speaker recognizability (defined as the degree by which listeners can recognize the converted voice as the target voice or discriminate between them).

There are several shortcomings in the previously published methodologies of evaluating VC systems. Often, distortion measures or statistical tests are used, which by themselves are inadequate for a signal that is meant to be heard by a human. When listening tests are conducted, they are typically small-scale and contain only a few source/target combinations. In addition, it is very difficult to compare results across different works, because every approach uses a different (often proprietary) speech database.

This work was supported by a grant from Intel Corporation and National Science Foundation grant IIS-9875950.

In this paper, we propose a more rigorous testing framework, in which we start with a corpus especially designed for the purposes of VC training. The design and creation of this corpus, planned for public release, is described in Section 2. In Section 3, we establish the degree by which listeners can distinguish different speakers in the corpus under various conditions, using a same/different sentence-pair listening test. The results are used to define a baseline against which conversion results can be compared. In Section 4, the conversion function responsible for mapping LPC parameters is tested. Using the listening test design of the previous section, subjects are asked to discriminate between converted, source, and target voices as LPC coded speech. In Section 5, we propose a novel way of constructing a converted utterance by mapping parameters of a spectral envelope and then predicting the residual from it. This is in contrast to other approaches which transform the residual. We present the algorithm and test it in our listening test framework.

2. SPEAKER DATABASE

The systematic training and evaluation of a VC system is facilitated by using a speech database that offers recordings of many different people producing the same sentences. Our goal was to design and develop a speech database that contained a "phonetically rich" set of sentences produced by multiple speakers. Additionally, the recording procedure was designed to result in a naturally good time-alignment between the same sentences of different speakers, which allows focusing on research of the segmental cues related to speaker identity as opposed to prosodic cues.

To select text that would provide coverage over the acoustical space, we ran a greedy algorithm on the list of sentences from the TIMIT and Harvard databases (a total of 1170 sentences). Our selection criterion maximized the number of occurrences of rare phonemes, while including as many unique diphones as possible. Using the top 50 sentences, each phoneme was represented on average 41 and at least 17 times; the number of unique diphones was 693.

First, a recording of a *template speaker* reading the selected sentences was made. Then, we invited 5 male and 5 female American English speakers, living in the US Pacific Northwest since childhood and between the ages of 21 and 29, as *corpus speakers*. During a recording session, they were asked to listen to a template sentence, whose text was also displayed on a screen, before being instructed to mimic the timing and accentuation pattern (but not pitch or voice quality) of that sentence on their own. The

speech and laryngograph signals were recorded at 22 kHz/16 bits in a professional sound-booth, using a high-quality headset condenser microphone. The entire database was force-aligned using the CSLU Speech Toolkit [1]. Time markers were created for each speech utterance at the beginning of a new HMM state (up to 3 per phoneme). The laryngograph track was processed to yield pitch mark estimates. Both the time markers and the pitchmarks were verified and corrected manually for a high degree of accuracy. An analysis of the time markers showed that the average sentence duration differences between mimicking speakers and the template speaker were less than 0.2%.

3. SPEAKER DISCRIMINATION BASELINE

In this section, we measure the degree by which listeners can distinguish different speakers of the corpus. The level of discrimination can be viewed as a baseline against which later VC results can be compared. Additionally, the database lends itself to discovering trends in the relative contributions of intonation, average fundamental frequency (F0), timing, and spectral detail to speaker recognizability. For these purposes, we designed a listening test with the following three conditions:

1. Natural: The unprocessed, original utterance, in which speaker-specific intonation and timing characteristics closely resemble those of the template speaker.
2. F0 and duration normalized: The durations and F0 of condition 1 were processed using PSOLA to yield a generic duration and F0 evolution. This was accomplished by "averaging" F0 and durations of each sentence from speakers of the same gender. Pitch, jitter, and duration information contributing to speaker discriminization is thus lost. Listeners can only make use of the short-term spectrum to discriminate speakers.
3. LPC spectrum: We performed a 22^{nd} order LPC analysis on condition 2 and synthesized utterances using only the LPC spectrum. Listeners must perform speaker discrimination based on the simplifying assumptions of the LPC model, one of the most significant of which is that the phase spectrum is minimum phase (except for unvoiced sounds where random phase was substituted). The speech has a coded quality.

A same/different task was chosen for the listening test, similar to [2]. In this type of task, listeners are played two different sentences and are asked whether they thought the sentences were spoken by either the same or by two different speakers. The listeners selected were completely unfamiliar with the speakers, since it is difficult to measure or control the degree of familiarity. Speaker *discrimination* by humans has been shown to be more accurate than speaker *recognition*, which is subject to memory limitations [3], and which can be significantly affected by the specific composition of a small speaker set [4]. Male and female voices were tested separately, because inter-sex confusions rarely occur (for example see [5]).

The following were additional design criteria. *Balance*: the stimuli were balanced in regards to gender, the number of "same" and "different" pairs, and the number of trials per condition. *Consistency*: the same speaker and sentence combinations were delivered for each listener. *Maximum variability*: a speaker never repeats the same sentence during the entire test. *Minimum bias*: the order of sentences, and the order of presentation of A and B within a trial were randomized. *Minimum learning of voices*: the order of gender presentation was switched from one to the other to slow down the learning of the voices. Also, the conditions were presented in sequence from 3 to 1 to delay the disclosure of full voice characteristics as much as possible.

Condition / Experiment	Males	Females
1: natural	84 (81-88)	95 (92-98)
2: F0 and duration normalized	83 (79-87)	89 (85-92)
3: LPC spectrum	71 (65-76)	88 (84-91)
LPC map	73 (69-77)	87 (83-90)
LPC map + residual prediction	74 (70-78)	84 (80-87)

Table 1. Results of the perceptual listening tests. Shown is the percent correct discrimination of speakers averaged over listeners and trials. The 95% confidence interval is in parentheses.

The average results of 16 listeners who each heard 120 sentence-pairs are displayed in Table 1. As expected, the average discrimination performance increased as more information was made available in the speech signal. The difference in discrimination between conditions 3 and 2 is significant for the set of male speakers, and between 2 and 1 for the set of female speakers ($\alpha = 0.01$). The significant increase in discrimination for males when adding information in the form of the complex LPC residual suggests that VC systems must be designed to be capable of producing spectral details not found in the LPC spectrum. This issue is addressed in Section 5.

4. LPC SPECTRUM MAPPING

There are two critical parts of a VC algorithm: the model and its parameters by which speech is modified to change the perceived speaker identity, and the function which is trained to predict target parameters from source parameters. The model and the conversion function have to be matched carefully to the task and available training data. The model must be effective in producing target speech naturally and accurately; at the same time the conversion function must be able to learn the source/target parameter relationship from the training data. In this section, we aim at evaluating the performance of the conversion function in isolation by listening to both natural and converted sentences as LPC-coded speech.

A number of VC algorithms in the literature have focused on converting the spectral envelope represented by a type of LPC parameter. In one of the earliest approaches, parameters were converted using a vector quantization (VQ) approach [6]. The discrete nature of this mapping was improved upon in [7] by using a Gaussian Mixture Model (GMM) within the framework of a sinusoidal synthesizer. In our earlier research, we mapped Bark-scaled Line Spectral Frequencies (LSF) by using a GMM that was estimated using a joint-density approach, creating the converted utterance by means of a residual LPC synthesizer [8].

During feature extraction, we first perform a pitch-synchronous sinusoidal analysis over 2 pitch periods. The discrete magnitude spectrum is upsampled and warped using the bark scale. Then, an application of the Levinson-Durbin algorithm on the autocorrelation sequence yields LPC filter coefficients [9]. It is important to note that a certain amount of excitation information is present in the LPC filter. Finally, the LPC filter parameters are converted to LSFs, which have more favorable interpolation properties. For the purpose of training, the features are time-aligned

with the aid of the time marker information and silences are removed. The final database, containing approximately 16,000 vectors for each speaker, is split into training (sentences 1-40) and test (sentences 41-50) sets.

The training and conversion procedures follow our previous work closely, please see [8] for details. First, the source and target vectors are joined to form a new vector space; a GMM of this space is estimated by the Expectation-Maximization (EM) algorithm, initialized by a generalized Lloyd algorithm [10]. After the log-likelihood stabilizes, a regression is performed which calculates the linear transformation components of the locally linear, probabilistic conversion function.

There are two free parameters in the training procedure, the number of mixture components Q and a scalar ε. The latter represents the magnitude of a perturbation added to the diagonal elements of the covariance matrices at each iterative estimation for the purpose of regularization. The choice of these parameters is problematic, because it is difficult to objectively measure speech quality and speaker accuracy. A spectral mean squared error (MSE) measure on the test set had a weak relationship to the desired outcome. For example, when Q was too high, the temporal evolution of the resulting spectra contained many discontinuities, even though the MSE was lower than compared to the results produced by a lower Q. An alternative measure is the signal to noise ratio (SNR), where the signal is defined as the converted sentence, and the error between it and the target sentence is defined as noise. This approach seemed to relate better with the desired speech quality of the output, as verified by an informal listening test. The final choices were $Q = 6$ and $\varepsilon = 0.0001$. The fact that the optimal number of components was so low (compared to earlier work) suggests that the mapping performed best when applying very broad transformations to the source vectors, which may indicate that there were not enough data to reliably estimate more components, and/or the data were "noisy" due to time-alignment problems during training.

During the conversion process, the source speech file is analyzed, its features transformed by the conversion function, and the target magnitude spectrum envelope is calculated by evaluating the predicted LPC system function, from which the converted speech is synthesized using a sinusoidal overlap/add system. The LPC system phase is used during voiced segments and random phase during unvoiced segments.

To evaluate the mapping performance, we conducted a listening test similar to Section 3, except we were now interested in the ability to discriminate between the converted speakers and their respective source and target speakers, which were reproduced by the same LPC synthesizer used during conversion. In this manner, listeners could only make use of differing feature information for discrimination. It should be noted that this test is very different from the conventional ABX test, which is based on *forced choice*: listeners are to decide whether the converted utterance is closer to the source or to the target speaker. The latter case does *not* imply that the converted speaker cannot be distinguished from the target speaker; thus, the speaker may in actuality not be recognizable.

Results from 12 listeners of the original listening group can be found in Table 1. The speaker distinction performance compares favorably to that of condition 3. This demonstrates that the conversion function is effective in producing a change in speaker. However, some degradation of the speech signal occurred, noticeable as a muffling effect.

5. RESIDUAL PREDICTION

Clearly, signal details beyond the LPC envelope contribute to the naturalness of speech and may also contain vital speaker information, apparent in the test results in Section 3 for the set of male speakers. To address this, several authors proposed ways of improving VC beyond changing the LPC spectral envelope by also changing the LPC residual. In [11], the authors formulated a codebook-based transformation of the source excitation characteristics by using a weighted combination of codeword filters, which were derived from the average source and target codeword excitation spectra. In [12], the excitation is modeled by a long delay neural net predictor whose parameters are mapped based on the maximum occurrence in a 2D histogram of vector correspondences. There exist also approaches based on estimating and modeling the glottal source, for example in [13] the voice type of a speaker can be converted between modal, breathy, or creaky.

In contrast to *transforming* an excitation waveform, we propose a method in which we *predict* the target residual from LPC parameters during voiced speech. The underlying assumption of this approach is that for a particular speaker and within some phonetically-similar class of voiced speech, the residuals are similar and predictable. Specifically, the residual's magnitude spectrum contains errors made by the spectral envelope fit (e.g. , zeros during a nasal), and the phase spectrum contains important information about the natural phase dispersion of the signal, as opposed to the minimum phase assumption of the LPC model. Another way of viewing this approach is as a speech coder with a speaker-dependent excitation codebook.

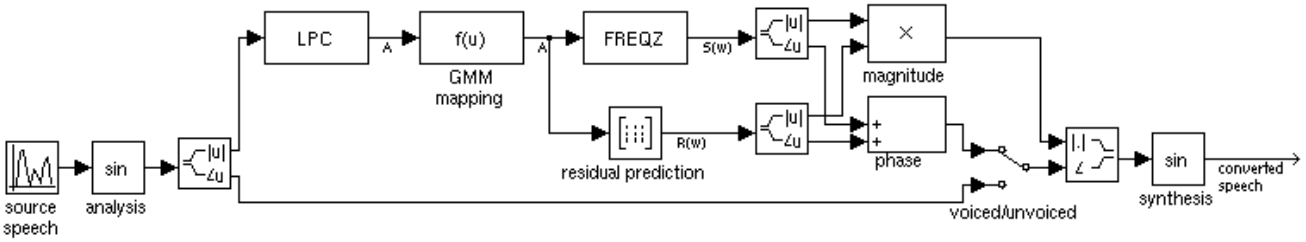
During training of the residual prediction module, a LPC cepstrum representation of all available voiced segments from the training set is clustered by a GMM with 32 mixture components. Each cepstral vector has a residual complex spectrum associated with it; the residual magnitude spectrum is calculated by subtracting the LPC log-magnitude envelope from the original log-magnitude spectrum, whereas the residual phase was given by the difference between the LPC system phase and the original phase spectrum. To make the codewords pitch-independent, the original residual vectors were upsampled to a common length using a nearest-neighbor interpolation scheme.

For each class, the residual codeword is calculated as follows: The magnitude spectrum is calculated by a weighted mean of all magnitude vectors, corresponding to the normalized probability of belonging to that class; the phase spectrum is set to the centroid phase. In the decoding stage, the posterior likelihood of an incoming cepstral vector is calculated and used as weights in predicting the residual magnitude by a weighted mean scheme from the residual codewords. The residual phase belonging to the class with maximum likelihood is chosen as the predicted phase. After this stage, the phases are unwrapped in time and smoothed by a 8-point FIR filter to reduce audible artifacts due to sudden changes in the residual phase. Finally, the residual spectrum is added to the LPC spectral envelope (see Figure 1).

The residual prediction module was tested by synthesizing sentences from their original spectral parameters only and was found to produce an output nearly indistinguishable from the original in informal listening tests.

We integrated the module with the spectral mapping system and also added a last stage in which the mean and variance of the source F0 is modified to match that of the target F0. The generated conversion sentences were compared to the original speech wave

Fig. 1.



Voice conversion algorithm based on converting the LPC spectrum and predicting the residual from the target LPC parameters.

files of the source and target speaker in a listening test in the same format as described earlier.

The results of the listening test in Table 1 show that the level of discrimination is significantly below that of the baseline for natural speech utterances within the speaker database. At first it may seem surprising that the level of discrimination dropped slightly as compared to the previous experiment. While it is true that the converted utterances contain more speaker information than before, they are also compared against natural waveforms, which in turn also contain many more speaker identity cues. The net effect is that the task had become more difficult. These results can be considered as a performance indicator of the "real world" task of mimicking another human with precision.

6. CONCLUSION

We have proposed a new VC algorithm based on predicting the LPC target residual from the target spectral envelope instead of transforming the source residual. To evaluate the level of accuracy by which the algorithm can convert voices such that they are indistinguishable from the target voice, a listening test was performed and the results compared with the appropriate baseline. The listening test was based on a same/different sentence-pair methodology, using combinations of 5 male and 5 female speakers. In another listening test, the conversion function implementing the transformation of LPC parameters was tested in isolation by listening to source, target, and converted voices as LPC coded speech.

The results show that a GMM can successfully transform the spectral envelope of a source speaker to be recognized as the spectral envelope of a target speaker. When comparing converted utterances with natural wavefiles, discrimination is still significantly lower than among natural samples. Additionally, the quality of converted speech is degraded. We speculate that improvements to the speech quality can be made in making changes to the manner in which the spectral mapping is trained to prevent problems that occur when the time-alignment is less than perfect. Another area of research is the optimal selection of mixture components and LPC order in the residual prediction module, as well as a more reliable way of extracting residual phase codewords.

A selection of the audio files used in this paper are available at http://cslu.cse.ogi.edu/tts.

7. REFERENCES

[1] John-Paul Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, Ph.D. thesis, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Beaverton, OR, USA, May 2000, Published as Technical Report CSE-00-TH-002.

[2] Astrid Schmidt-Nielsen and Derek P. Brock, "Speaker recognizability testing for voice coders," in *Proceedings of ICASSP '96*, Atlanta, GA, May 1996, vol. 2, pp. 1149–1152.

[3] Jody Kreiman and George Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, pp. 265–275, 1991.

[4] Astrid Schmidt-Nielsen and K. R. Stern, "Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection," *Journal of the Acoustical Society of America*, vol. 79, no. 4, pp. 1174–1177, April 1986.

[5] Panos E. Papamichalis and George R. Doddington, "A speaker recognizibility test," in *Proceedings of ICASSP '84*, San Diego, California, 1984, vol. 2, p. 18B.6.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of ICASSP '88*, New York, NY, 1988, IEEE, pp. 655–658.

[7] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," in *Proceedings of Eurospeech '95*, Madrid, Spain, September 1995, pp. 447–450.

[8] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1998, vol. 1, pp. 285–288.

[9] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 4, pp. 121–173. Elsevier Science, Amsterdam, Holland, 1995.

[10] Nanda Kambhatla, *Locally Linear Models for Statistical Data Processing*, Ph.D. thesis, Oregon Graduate Institute, Dec. 1995.

[11] Levent M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[12] K. S. Lee, D. H. Youn, and I. W. Cha, "A new voice transformation method based on both linear and nonlinear prediction analysis," in *Proceedings of ICSLP '96*, Philadelphia, PA, October 1996, vol. 3, pp. 1401–1404.

[13] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, February 1995.