SMOOTH GMM BASED MULTI-TALKER SPECTRAL CONVERSION FOR SPECTRALLY DEGRADED SPEECH

Chuping Liu[1], Qian-Jie Fu[2,3], and Shrikanth S. Narayanan[1]

[1]Department of Electrical Engineering, [2]Department of Biomedical Engineering,
University of Southern California, Los Angeles, CA, 90007, USA
[3]Department of Auditory Implants and Perception, House Ear Institute
2100 West Third Street, Los Angeles, CA, 90057, USA
Email: chupingl@usc.edu

ABSTRACT

Because of the limited spectro-temporal resolution associated with the implant device, cochlear implant (CI) patients are more susceptible to talker variability than normal hearing (NH) listeners. In the present study, the effect of a smooth GMM based spectral conversion algorithm on multi-talker sentence recognition was tested in CI patients. In a model of CI speech processing (4-16 channels of spectrally degraded speech), talker distortion was significantly reduced with relatively few (~64) GMM components. CI patients' sentence recognition was measured for one male (M1) and one female (F1) talker, as well as for spectrally converted speech (from M1 to F1 and from F1 to M1). Overall, CI users were sensitive to talker differences; some subjects performed better with M1, others with F1. After converting the spectrum of the less-understood talker to that of the better-understood talker, recognition of the less-understood talker's speech was significantly improved. The results suggest that smooth GMM-based spectral conversion may improve CI patients' multi-talker speech recognition.

1. INTRODUCTION

Normal hearing (NH) listeners are able to understand speech from a variety of talkers, despite differences in acoustic characteristics (e.g., gender, age, accent, etc.). NH listeners are thought to use some form of "speaker normalization" to process speech from multiple talkers, in which the speech patterns from a variety of talkers are normalized to a central pattern template [1]. However, talker variability has been shown to affect multi-talker speech recognition by NH listeners. Multi-talker speech recognition is even more difficult for hearing-impaired (HI) and cochlear implant (CI) listeners [2]. For CI patients, increased susceptibility to talker variability may be due to the limited spectro-temporal cues needed to perform perceptual normalization. Distorted spectral information, due to the mismatch between the input acoustic frequency and electrode place of stimulation, may also contribute to poorer multi-talker speech recognition.

In a previous study [3], a speaker normalization algorithm was used to improve multi-talker Chinese vowel recognition in a 4-channel acoustic simulation of CI processing. The analysis filter bank was adjusted to match the optimal reference pattern according to the ratio of the mean third formant frequency (F3) values between each talker in the multi-talker set and the reference talker (the talker that produced the best vowel recognition for each subject). Results indicated a small but significant improvement in subjects' overall multi-talker vowel recognition; larger improvements were observed for each subject's least-understood talker. However, linear warping according to talkers' F3 position may not be sufficient to effectively normalize acoustic differences between talkers; other, more complex acoustic differences (e.g., formant bandwidth, spectral envelope, spectral tilt, prosodic variations, etc.) may also require normalization. By using a continuous statistical model (e.g., Gaussian mixture model, or GMM) to compensate for acoustic differences between talkers, speech from a variety of talkers may be better matched to CI users' optimal speech patterns.

In the present study, a smooth GMM-based spectral conversion algorithm was used to convert speech patterns between a male and a female talker. A CI simulation model was used to test the spectral conversion effect under different spectral degradation. Sentence recognition with/without algorithm was tested in five CI patients. It was hypothesized that performance would be improved for the less-understood talker with the spectral conversion algorithm.

2. METHODS

2.1. GMM-based spectral conversion

A GMM allows the distribution of the observed parameters, represented by m mixture Gaussian components in the form of

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}, \mu_i, \Sigma_i)$$

where α_i denotes the priority probability of component i ($\sum_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$) and $N(\mathbf{x}, \mu_i, \Sigma_i)$ denotes the normal distribution of the i^{th} component with mean vector μ_i and covariance matrix Σ_i in the form of

$$N(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \Sigma_i^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)]$$

where p is the number of vector dimensions. The parameters of the model (α, μ, Σ) can be estimated using the well-known expectation maximization (EM) algorithm [4].

Let $\mathbf{x} = [\mathbf{x}_1 \ \mathbf{x}_2 \ ... \ \mathbf{x}_n]$ be the sequence of n spectral vectors produced by the source talker, and let $\mathbf{y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ ... \ \mathbf{y}_n]$ be the time-aligned spectral vectors produced by the target talker. The objective of the proposed spectral conversion algorithm is to define a conversion function $F(x_t)$ such that the total conversion error of spectral vectors

$$\varepsilon = \sum_{t=1}^{n} (\mathbf{y}_t - F(\mathbf{x}_t))^2$$

is minimized over the entire acoustic space, using GMM. A minimum mean square error (MMSE) method was used to estimate the conversion function after GMM modeling of the source talker's spectral distribution [5,6,7]. The conversion function was:

$$F(\mathbf{x}_t) = \sum_{i=1}^{m} P(C_i \mid \mathbf{x}_t)[\mathbf{v}_i + \mathbf{T}_i \Sigma_i^{-1} (\mathbf{x}_t - \mu_i)]$$

where $P(C_i \mid \mathbf{x}_t)$ is the posterior probability that the i^{th} Gaussian component generates $\mathbf{x}_t$; $\mathbf{v}_i$ and $\mathbf{T}_i$ are the mean target vector and cross-covariance matrix of the source and target vectors, respectively. When a diagonal conversion is used (i.e., $\mathbf{T}_i$ and $\sum_i$ are diagonal), the above optimization problem is thus split into a scalar optimization problem.

2.2. Speech analysis and synthesis

A Mel-scaled LSF feature was used for speech analysis because it is perceptually based and has smooth interpolation characteristics [6]. After frame-based speech analysis (20ms frame length; 7ms shift interval) and LPC coefficients extraction, LPC spectrum was converted to Mel-warped spectrum according to the relationship $M(f) = 1125 \ln(1 + f/700)$ [4]. The warped spectrum was then uniformly re-sampled with spline cubic phase interpolation to obtain the Mel-scaled LPC spectrum. A least square fit was used to convert the Mel-scaled LPC spectrum to Mel-scaled LPC coefficients, which were then converted to Mel-scaled LSF coefficients.

In speech synthesis, a source filter model was applied in which the filter was the converted LPC spectrum and the source was the source talker's residual.

2.3 CI speech simulation

A noise-band vocoder was used to simulate a CI speech processor fitted with the Continuous Interleaved Sampling (CIS) strategy [8]. The processor was implemented as follows. The signal was first pre-emphasized with a filter coefficient of 0.95. The input frequency range (100 – 6000 Hz) was band-passed into a number of frequency analysis bands (24 dB/octave filter slope), distributed according to Greenwood's formula [10]. The temporal envelope was extracted from each frequency band by half-wave rectification and low-pass filtering (160 Hz envelope filter). The envelope of each band was used to modulate a wideband noise, which was then spectrally limited by the same band-pass filter used for frequency analysis. Finally, the modulated carriers of each band were summed and the overall level was adjusted to be the same RMS level as the original speech.

2.4 Implementation framework of spectral conversion algorithm with CI simulation

Figure 1 illustrates the GMM-based spectral conversion algorithm used with the CI simulation. The major three components of the model (i.e., GMM-based spectral conversion, speech analysis/synthesis and CI simulation) are detailed in Sections 2.1 - 2.3.

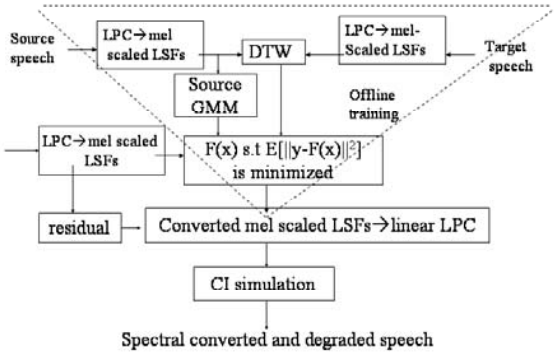

Figure 1: Implementation framework of GMM-based spectral conversion for spectrally degraded speech.

The spectral conversion algorithm was tested using IEEE sentences [11] recorded with one male (M1) and one female (F1) talker. The training dataset included 100 sentences randomly selected from the database, which resulted in over 60,000 Mel-scaled 25th order LSF feature vectors to train the GMM. To reduce computational complexity, only diagonal spectral conversion was tested. The testing dataset included the entire database.

4. RESULTS AND DISCUSSION

4.1. Objective test

The number of GMM components is an important design factor to consider. The degree of spectral conversion was calculated according to the average MFCC distortion between the converted speech and target speech in the training set, which was then normalized by the acoustic distortion between source speech and target speech. As shown in Figure 2, acoustic distortion decreases with increasing GMM components. However, acoustic distortion only marginally decreases for GMM components numbering more than 16, saturating at -4dB, implying that relatively few GMM components are needed for spectral conversion. Hence, in the present study, the number of GMM components was limited to 64.

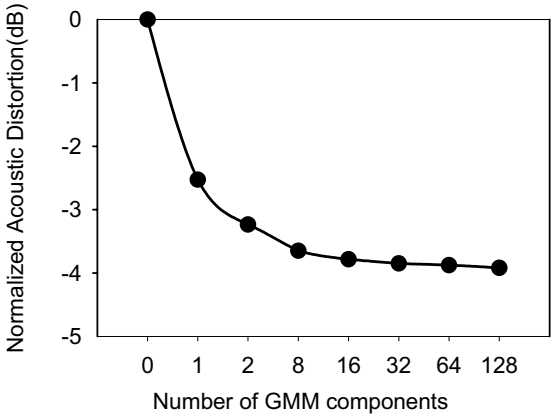

Figure 2: Normalized acoustic distortion between converted speech and target speech as a function of the number of GMM components.

A typical spectral envelope conversion is shown in Figure 3. The target envelope is an approximation due to the possible errors in dynamic time warping and statistical modeling of the source vectors. The spectral conversion is shown to transform the formant position/bandwidth, spectral tilt and energy distribution of the source spectrum toward that of the target.

Variation in spectral envelope between talkers is thought to relate to vocal production anatomy. Vocal Tract Length (VTL) is a primary cue for talker classification because it provides strong acoustic correlates to talker gender and talker identity [9]. In order to compare the

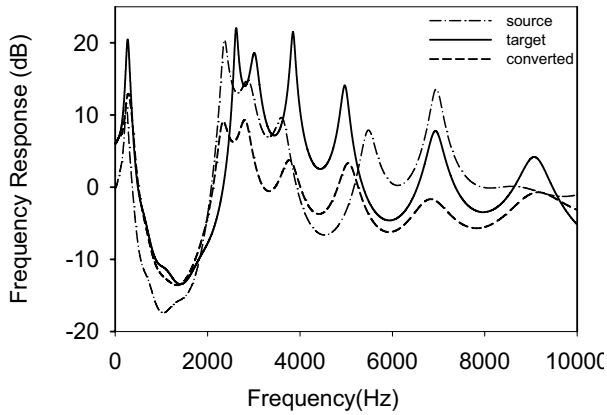

Figure 3. Spectral envelope conversion (64 GMM components with diagonal conversion). Dash-dotted line: source envelope. Solid line: approximated target envelope. Dashed line: converted envelope.

degree of spectral conversion, the VTLs of talkers M1 and F1 were compared to spectrally converted VTLs. VTLs were estimated from formant frequency measurements for the three /ɛ/s in the sentence: "It's easy to tell the depth of the well." As shown in Table 1, the GMM-based spectral conversion effectively transformed the VTL of the source talker toward that of the target talker.

Table 1: Estimated VTLs for source and target talkers.

	F1	M1	F1→ M1	M1→ F1
VTL (cm)	14.75	16.43	15.14	15.51

The degree of spectral conversion shown in Figure 3 and Table 1 were calculated using the speech tokens without spectral degradation. To see the degree of conversion for spectrally-degraded speech (as is typically experienced by CI users), the GMM-based algorithm was tested using the CI simulation with 4–16 spectral channels. Figure 4 shows the acoustic distortion between converted speech and target speech (normalized to the acoustic distortion between source and target talkers under original speech condition), as a function of the number of spectral channels. Talker distortion decreased similarly (-0.42dB/channel) with or without spectral conversion, as the number of spectral channels was reduced. Talker distortion was fairly uniformly reduced with an average of –2.73dB across all spectral resolution condition and is significant (paired t-test: $p < 10^{-5}$) with conversion algorithm.

4.2. Formal listening test with CI users

The GMM-based spectral conversion algorithm was also assessed in 5 CI patients. IEEE sentence recognition in quiet was measured for 4 talker conditions: M1, F1, M1→F1 and F1→M1. Sentence recognition was measured in free field using subjects' clinically assigned speech processors. Overall, talker preference significantly affected

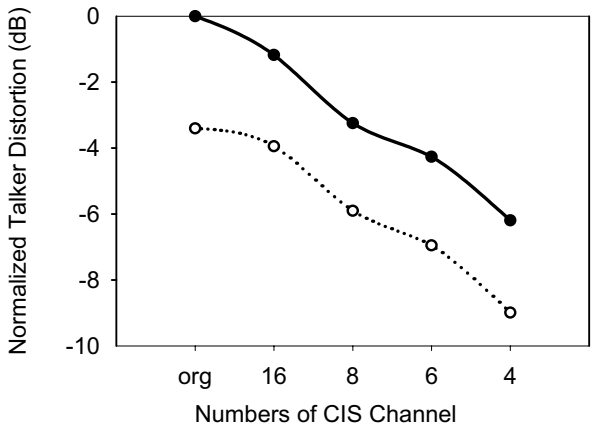

Figure 4: Normalized talker distortion as a function of number of channels. Solid line: without spectral conversion. Dotted line: with spectral conversion.

performance (paired t-test: p=0.01); however, some performed better with M1, others with F1. Figure 5 shows the net perception rate change with better-understood/less-understood talker as the spectral conversion target.

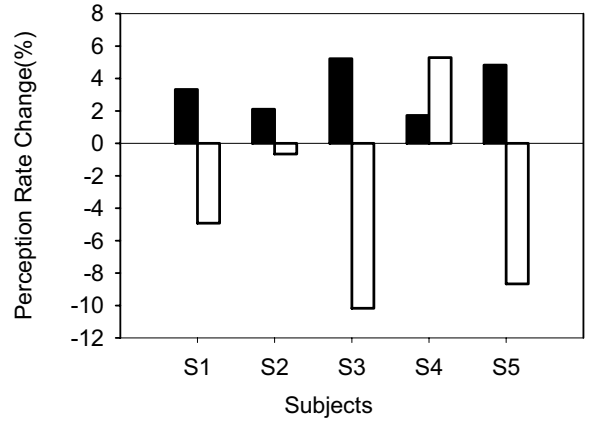

Figure 5: Perception rate change with spectral conversion. Black bar: better-understood talker is the target. White bar: less-understood talker is the target.

It is observed that large increase and decrease of performance occurred in the subjects S3 and S5, implying that these two subjects were more susceptible to multi-talker variability than S1 and S2. Statistical analysis revealed that although performance was not significantly decreased when less-understood talker was the target due to S4, recognition performance was significantly improved when better-understood talker was the target (p=0.004). However, it is noted that subject S4 had increased performance in both directions of conversion. For this individual, spectral conversion had resulted in creating an optimal speech pattern somewhere between better-understood talker and less-understood talker. This paradoxical result for S4 may be an artifact of the low number of talkers used in the experiment.

5. CONCLUSIONS

This paper presented and evaluated a GMM-based spectral conversion algorithm under spectrally degraded speech, motivated by CI users' needs. In a model of CI speech processing (4- to 16-channel noise vocoder), talker distortion was significantly reduced with the algorithm, using relatively few GMM components (64); VTLs for spectrally converted talkers approached those of the target talkers. Sentence recognition performance was significantly improved for the less-understood talker with the spectral conversion algorithm. These results suggest GMM-based spectral conversion may enhance CI users' multi-talker speech recognition by transforming the acoustic characteristics of multiple talkers toward the optimal spectral representation for individual CI patients.

6. REFERENCES

[1] D.B. Pisoni, "Long term memory in speech perception: some new findings on talker variability, speaking rate, and perceptual learning", Speech Communication, vol.13, no.1-2, pp.109-125, 1993.

[2] J.W. Mullennix, D.B. Pisoni, and C.S. Martin, "Some effects of talker variability on spoken word recognition", J.Acoust.Soc.Am., vol. 85, no.1, pp. 365-378, 1989.

[3] X. Luo and Q.-J. Fu, "Speaker Normalization for Chinese Vowel Recognition in Cochlear Implants," IEEE Transaction on Biomedical Engineering, vol.52, no. 7, pp. 1358 – 1361, 2005.

[4] X.-D. Huang, A. Acero, and H.-W. Hon, "Spoken language processing- a guide to theory, algorithm, and system development", Prentice Hall, 2001.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Transaction on Speech and Audio Processing, vol.6, no. 2, 1998.

[6] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-Speech synthesis," ICASSP, pp.285-288, 1998.

[7] J.M. Mendel, "Lessons on estimation theory for signal processing,communications and control," Prentice Hall, 1995.

[8] B.S. Wilson, C.C. Finley, D.T. Lawson, R.D. Wolford, D.K. Eddington, and W.M. Rabinowitz, "New levels of speech recognition with cochlear implants," Nature 352, pp. 236-238, 1991.

[9] J.A. Bachorowski and M.J. Owren, "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J.Acoust.Soc.Am., vol. 106, no.2, pp. 1054-1063, 1999.

[10] D.D. Greenwood, "A cochlear frequency-position function for several species – 29 years later," J.Acoust.Soc.Am., vol. 87, no.2, pp. 2592-2605, 1990.

[11] IEEE, "IEEE recommended practice for speech quality measurements," Institute of Electrical and Electronic Engineers, New York, 1969.