 [DOI: 10.1121/1.3062145] Published Online 21 January 2009

Effect of bandwidth extension to telephone speech recognition in cochlear implant users

Chuping Liu
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089
chupingl@usc.edu

Qian-Jie Fu
Department of Biomedical Engineering, University of Southern California, Los Angeles, California 90089
and Department of Auditory Implants and Perception, House Ear Institute,
2100 West Third Street, Los Angeles, California 90057
qfu@hei.org

Shrikanth S. Narayanan
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089
shri@sipi.usc.edu

Abstract: The present study investigated a bandwidth extension method to enhance telephone speech understanding for cochlear implant (CI) users. The acoustic information above telephone speech transmission range (i.e., 3400 Hz) was estimated based on trained models describing the relation between narrow-band and wide-band speech. The effect of the bandwidth extension method was evaluated with IEEE sentence recognition tests in seven CI users. Results showed a relatively modest but significant improvement in the speech recognition with the proposed method. The effect of bandwidth extension method was also observed to be highly dependent on individual CI users.

PACS numbers: 43.71.Ky, 43.64.Me, 43.60.Dh [DS]
Date Received: July 31, 2008 **Date Accepted:** December 8, 2008

1. Introduction

Telephone use is still challenging for many deaf or hearing-impaired individuals including cochlear implant (CI) users. According to a previous study (Kepler *et al.*, 1992), there are three major contributors to the difficulties in telephone communication: the limited frequency range, the elimination of visual cues, and the reduced audibility of telephone signal. For example, the telephone bandwidth in use today is limited to 300–3400 Hz. Compared to speech in face-to-face conversational settings, telephone speech does not convey information above 3400 Hz, which is useful in the identification of many speech sounds, notably certain consonants such as fricatives. Since CI users generally receive frequency information up to approximate 8 kHz or even higher, the narrow-band telephone speech may present an obstacle even when they can achieve a fairly good wide-band speech perception.

Previous studies have assessed the capability of CI patients to communicate over telephones. While many CI patients were capable of certain degree of communication over the phones, speech understanding was significantly worse than with broad-band speech (Milchard and Cullington, 2004; Ito *et al.*, 1999; Fu and Galvin, 2006). For example, word discrimination score obtained from telephone speech was decreased by 17.7% than those with wide-band speech. Analysis of the word errors revealed that the place of articulation was the predominant type of error (Milchard and Cullington, 2004). On the other hand, investigation of telephone use among CI recipients reported that 70% of the respondents communicated via the telephone, of which 30% used cellular phones (Cray *et al.*, 2004). Hence, improved capability to understand telephone speech using just auditory cues will increase the opportunities for the use of the

telephone and will promote independent living, employment, socialization, and self-esteem in CI users.

To improve the telephone communication ability of hearing-impaired people, one solution, albeit expensive, is to change the current public switched telephone network to transmit wide-band speech and to enrich the spoken information with videos. This is, however, difficult to accomplish in the near future. A more economical and near term approach is to add external equipment to enhance the audibility of telephone speech. For example, the telephone adapter, which was used to reduce noise level in the telephone and to record telephone speech into a tape recorder, was found to boost speech-tracking scores in CI users (Ito *et al.*, 1999). Yet, such auxiliary instruments may not be easy to obtain, especially in mobile communication. Another potential approach is to improve speech processing and transmission technique. A previous study (Terry *et al.*, 1992) investigated frequency-selective amplification and compression via digital signal processing techniques to compensate for high-frequency hearing loss in hearing-impaired people. Nevertheless, the approach required audiometric data from individual users to achieve the best performance.

On the other hand, to overcome the deficit of telephone speech in terms of narrow bandwidth, bandwidth extension as a front end processing was studied (e.g., Nilsson and Kleijn, 2001; Jax and Vary, 2003). For example, Jax and Vary (2003) proposed an approach to extend telephone bandwidth to 7 kHz based on hidden Markov model. Nilsson and Kleijn (2001) studied a bandwidth extension approach to avoid overestimation of high-band energy. Through listening tests, the method was shown to reduce the degree of artifacts. Yet, it is not clear how much gain the bandwidth-extension method can actually bring to speech recognition with listeners, especially for CI users.

In this study, we propose a bandwidth-extension method to enhance telephone speech. Gaussian mixture model (GMM) was used to model the spectrum distribution of narrow-band speech. The relationship between wide-band and narrow-band speech was learned a priori in a data driven fashion and was used to recover the missing information based on the available telephone band speech. Such an approach does not require auxiliary instruments and patient data for its implementation. We then studied the effect of the proposed bandwidth-extension method on speech recognition performance in CI users.

2. Methods

The step to expanding narrow-band speech to wide-band speech basically consists of two parts: spectral envelope extension and excitation spectrum extension, which are introduced in Secs. 2.1 and 2.2, respectively.

2.1 GMM-based spectral envelope extension

A GMM represents the distribution of the observed parameters by m mixture Gaussian components in the form of

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

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

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

where p is the vector dimension. The parameters of the model (α, μ, Σ) can be estimated using the well-known expectation maximization algorithm.

Let $\mathbf{x}=[\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_n]$ be the sequence of n spectral vectors produced by the narrow-band telephone speech, and let $\mathbf{y}=[\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_n]$ be the time-aligned spectral vectors produced by the wide-band speech. The objective of the bandwidth-extension method was to define a conversion function $F(x_t)$ such that the total conversion error of spectral vectors

$$\varepsilon = \sum_{t=1}^{n} (\mathbf{y}_t - F(\mathbf{x}_t))^2 \tag{3}$$

was minimized over the entire training spectral feature set, using the trained GMM that represents the feature distribution of the telephone speech. A minimum mean square error method was used to estimate the conversion function. The conversion function was (Stylianou *et al.*, 1998; Kain and Macon, 1998)

$$F(\mathbf{x}_t) = \sum_{i=1}^{m} P(C_i|\mathbf{x}_t)[\mathbf{v}_i + \mathbf{T}_i\Sigma_i^{-1}(\mathbf{x}_t - \mu_i)], \tag{4}$$

where $P(C_i|\mathbf{x}_t)$ is the posterior probability that the ith Gaussian component generates $\mathbf{x}_t$; $\mathbf{v}_i$ and $\mathbf{T}_i$ are the mean wide-band spectral vector and the cross-covariance matrix of the wide-band and narrow-band spectral vectors, respectively. When a diagonal conversion is used (i.e., $\mathbf{T}_i$ and Σ_i are diagonal), the above optimization problem simplifies into a scalar optimization problem and the computation cost is greatly decreased.

2.2 Excitation spectrum extension

Two methods are considered for excitation spectrum extension in this study (Makhoul and Berouti, 1979): spectral folding and spectral translation. Spectral folding simply generates a mirror image of the narrow-band spectrum for high-band spectrum. The implementation of spectral mirroring was equivalent to upsampling the excitation signal in the time domain by zero padding. This almost added no extra cost in the processing. Yet, the energy in the reconstructed high band is typically overestimated with this approach; the harmonic pattern of the restored high band is a flipped version of the original narrow-band spectrum, centered around the highest frequency of the narrow-band speech. Spectral translation, on the other hand, did not have these problems, but involves more expensive computation. The excitation spectrum of the narrow-band speech, obtained from Fourier transformation of the time domain signal, is translated to the high-frequency part and padded to fill the desired whole band. A low pass filter is applied to do spectral whitening, such that the discontinuities between the translations are smoothed. The extended wide-band excitation in the time domain is then obtained from inverse Fourier transformation.

2.3 Speech analysis and synthesis

In this study, Mel-scaled line spectral frequency (LSF) features (18th order) and energy were extracted to model the spectral characteristics of speech in a 19 dimensional space. The spectral features between narrow-band and wide-band speech were aligned with dynamic time warping computation. The spectral mapping function between narrow-band and wide-band speech was trained with 200 randomly selected sentences from the IEEE database (100 sentences from a female talker and the other 100 sentences from a male talker). The excitation component between 1 and 3 kHz was used to construct the high-band excitation component because the spectrum in this range was relatively white. A low pass Butterworth filter (first order with cutoff frequency 3000 Hz) was used to do spectral whitening. The synthesized high-band speech (i.e., frequency information above 3400 Hz) was obtained from high pass filtering the convolution result of the extended excitation and extended spectrum. It was then appended to the original telephone speech to render the reconstructed wide-band speech that covered the frequency band from 300 to 8000 Hz.

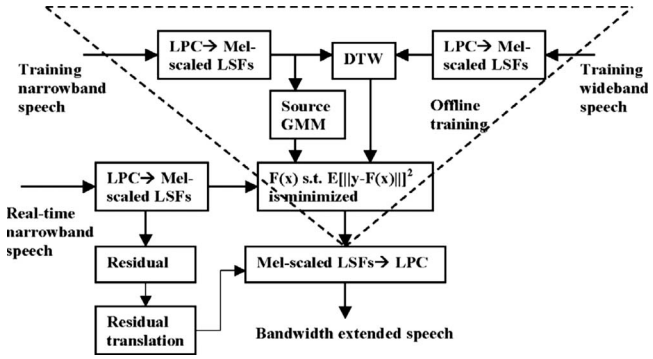

Fig. 1. Implementation framework of the GMM-based bandwidth-extension method.

2.4 Implementation framework of the bandwidth-extension method

Figure 1 illustrates the GMM-based bandwidth-extension method. The three major components of the model (i.e., GMM-based spectral envelope extension, excitation spectrum extension, and speech analysis/synthesis) are as detailed in Secs. 2.1 and 2.3.

2.5 Test materials and procedures

The test materials in this study were IEEE (1969) sentences, recorded from one male talker and one female talker at the House Ear Institute with a sampling rate of 22 050 Hz. The narrowband telephone speech was obtained by bandpass filtering the above wide-band speech (ninth order Butterworth filter, bandpass between 300 and 3400 Hz) and was downsampled to 8 kHz. Three conditions were tested: restored wide-band speech (carrying information up to 8 kHz), telephone speech (carrying information up to 3.4 kHz), and originally recorded wide-band speech (carrying information up to 11 kHz). All sentences were normalized to have the same long-term root mean square value. Note that the GMM training sentences (i.e., 200 randomly selected sentences) were also bandwidth extended and included in the listening test to increase the available speech materials for the experiment.

Seven CI subjects (two women and five men) participated in this study. Table 1 lists relevant demographics for the CI subjects. All subjects were native speakers of American English and had extensive experience in speech, recognition experiments. For all the listening conditions including restored wide-band speech, telephone speech, and originally recorded wide-band speech, subjects were tested using their clinically assigned speech processor and

Table 1. Subject demographics for the CI patients who participated in the present study.

Subject	Age	Gender	Etiology	Implant type	Strategy	Duration of implant use (years)
S1	55	M	Hereditary	Freedom	ACE	1
S2	62	F	Genetic	Nucleus-24	ACE	2
S3	48	M	Trauma	Nucleus-22	SPEAK	13
S4	67	M	Hereditary	Nucleus-22	SPEAK	14
S5	64	M	Trauma/unknown	Nucleus-22	SPEAK	15
S6	75	M	Noise induced	Nucleus-22	SPEAK	9
S7	72	F	Unknown	Nucleus-24	ACE	5

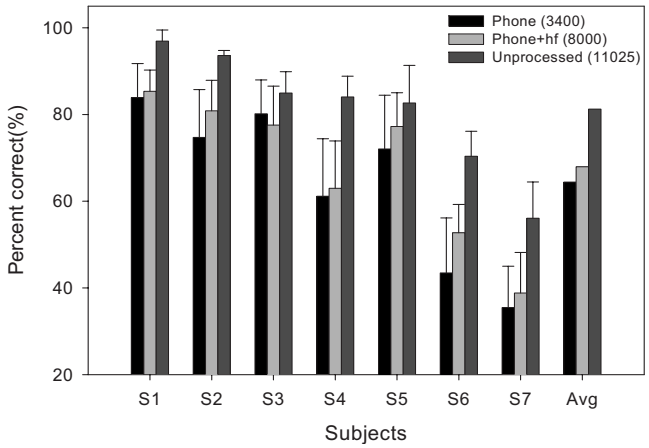

Fig. 2. Sentence recognition performance for individual CI subjects with and without the bandwidth-extension method, and with the unprocessed wide-band speech. The error bars indicate one standard deviation.

comfortable volume/sensitivity settings. As shown in Table 1, the subjects used ACE (Skinner *et al.*, 2002) or SPEAK strategy (Seligman and McDermott, 1995). The maximum number of activated electrodes is typically 6 for SPEAK strategy and 8 for ACE strategy, respectively. While the number of activated electrodes is the same for both telephone speech and broad-band speech, the number of total usable electrodes is different. In general, all 20 electrodes will be used when listening to broad-band speech while only 13 electrodes will be used for telephone speech. Once testing began, these settings were not changed. Subjects were tested while seated in a double-walled sound-treated booth (IAC). Stimuli were presented via a single loud speaker at 65 dBA. The test order of different conditions was randomized for each subject. No feedback was provided during the test.

3. Results and discussion

The sentence recognition performance with and without the restored high-band components is shown in Fig. 2, together with the performance with the naturally recorded wide-band speech. Note that the subjects are ordered according to their performance with wide-band speech. On average, compared to the performance with the naturally recorded wide-band speech, the performance with the narrow-band telephone speech was about 16.8% lower, which was significant (paired t-test: $p \leqslant 0.001$). The recognition score with the bandwidth-extension method was about 3.5% higher than without the bandwidth-extension method. The improvement was small but significant (paired t-test, $p = 0.050$). Yet, the performance with the bandwidth-extension method was still significantly lower than with the unprocessed wide-band speech (paired t-test, $p \leqslant 0.001$).

Figure 2 demonstrates substantial cross subject variability in performance. First, the cross subject variability was observed in terms of the performance for the same test materials. For example, subject S1 obtained over 80% correct under with and without the restored high-band component conditions. In contrast, subject S7 obtained only about 40% in average. Second, the cross subject variability was observed in terms of the effect of the bandwidth-extension method. For example, subject S6 achieved about 10% improvement with the restored high-band information; while subject S3 had even about 3% deficit in performance with the restored high-band information.

4. Discussion

The present study showed a 16.8% performance drop in CI users' listening to narrow-band telephone speech than listening to the originally recorded wide-band speech. This percentage drop was similar to the performance drop reported in Milchard and Cullington, 2004, although

the testing materials and testing procedures were different between these two studies. In the current study, seven CI subjects were tested with IEEE (1969) sentences. In Milchard and Cullington's (2004) study, ten CI subjects were tested with 80 consonant-vowel-consonant type stimuli (e.g., BAD BAG BAT BACK) using the four alternative auditory feature test procedure. The present study confirmed the findings in previous studies that the bandwidth effect was substantial in CI listeners.

The observed cross subject performance difference may be due to different CI device settings and different electropsychoacoustic listening patterns across subjects. For example, for those CI users whose speech processor encoded more information on the high-band speech, the potential benefit of the bandwidth-extension method may be relatively larger than the other CI users.

In the present study, a bandwidth-extension method was proposed to improve the telephone speech recognition performance in CI listeners. Although speech recognition was improved with the proposed bandwidth-extension method in a significant manner, the improvement was relatively small compared to the observed 16.8% performance drop from wide-band speech to telephone speech. There are four possible reasons for this marginal improvement. First, the proposed bandwidth-extension method only recovered information up to 8 kHz, while the 16.8% performance drop was the performance difference between wide-band speech (11 kHz) and narrow-band telephone speech (3.4 kHz). It was not clear how much the recognition benefit might be for the acoustic information between 8 and 11 kHz. Second, in this study, Mel-scaled LSF features were used, which placed lower resolution on the high-frequency components. The feature order used for speech analysis was the same (18th order) for both wide-band and narrow-band speech, although their frequency ranges were different. Such signal processing procedures may not result in high accuracy in parameter estimation. Third, due to the nature of speech synthesis, it was difficult to accomplish a synthesis without perceptual distortion. The introduced artifacts may be very detrimental for CI listeners, who typically receive degraded spectrotemporal information. Finally, performance with the bandwidth-extended speech was acutely measured in CI listeners in free field; the potential benefit with the bandwidth extended method might be underestimated since the training effect was not taken into account.

5. Conclusions

This paper studied a bandwidth-extension method to enhance telephone speech understanding in CI users. The lost high-band acoustic information was estimated based on the available narrow-band telephone speech and a pretrained relation between narrow-band and wide-band speech. The narrow-band excitation was extended to wide-band excitation by spectral translation. A source filter model was used to synthesize estimated wide-band speech, whose high-band frequency information was filtered out and appended to the original telephone speech. The effect of bandwidth-extension method was evaluated with IEEE (1969) sentence recognition tests in seven CI users. Results showed that CI speech recognition was significantly improved with the bandwidth-extension method, although it was relatively small compared to the performance drop seen from the wide-band speech to telephone speech. The benefit of the bandwidth-extension method was also highly dependent on individual CI users.

Acknowledgments

We acknowledge all the subjects that participated in this study. Research was supported in part by NIH-NIDCD.

References and links

Cray, J. W., Allen, R. L., Stuart, A., Hudson, S., Layman, E., and Givens, G. D. (**2004**). "An investigation of telephone use among cochlear implant recipients," Am. J. of Audiology **13**, 200–212.

Fu, Q. J., and Galvin, J. J. (**2006**). "Recognition of simulated telephone speech by cochlear implant users," Am J. Audiol. **15**, 127–32.

IEEE (**1969**). *IEEE Recommended Practice for Speech Quality Measurements* (Institute of Electrical and Electronic Engineers, New York).

Ito, J., Nakatake, M., and Fujita, S. (**1999**). "Hearing ability by telephone of patients with cochlear implants," Otolaryngol.-Head Neck Surg. **121**, 802–804.
Jax, P., and Vary, P. (**2003**). "On artificial bandwidth extension of telephone speech," Signal Process. **83**, 1707–1719.
Kain, A., and Macon, M. W. (**1998**). "Spectral voice conversion for text-to-speech synthesis," IEEE ICASSP, pp. 285–288.
Kepler, L. J., Terry, M., and Sweetman, R. H. (**1992**). "Telephone usage in the hearing-impaired population," Ear Hear., **13**, 311–319.
Makhoul, J., and Berouti, M. (**1979**). "*High-frequency regeneration in speech coding systems*," IEEE ICASSP, pp. 428–431.
Milchard, A. J., and Cullington, H. E. (**2004**). "An investigation into the effect of limiting the frequency bandwidth of speech on speech recognition in adult cochlear implant users." Int. J. Audiol., **43**, 356–362.
Nilsson, M., and Kleijn, W. B. (**2001**). "Avoiding over-estimation in bandwidth extension of telephony speech," IEEE ICASSP, pp. 869–872.
Seligman, P. M., and McDermott, H. J. (**1995**). "Architecture of the spectra-22 speech processor," Ann. Otol. Rhinol. Laryngol. Suppl. **166**, 139–141.
Skinner, M. W., Arndt, P. L., and Staller, S. J. (**2002**). "Nucleus 24 advanced encoder conversion study: Performance versus preference," Ear Hear. **23**, 2S–17S.
Stylianou, Y., Cappe, O., and Moulines, E. (**1998**). "Continuous probabilistic transform for voice conversion," IEEE Trans. Commun. **6**, 131–142.
Terry, M., Bright, K., Durian, M., Kepler, L., Sweetman, R., and Grim, M. (**1992**). "Processing the telephone speech signal for the hearing impaired," Ear Hear. **13**, 70–79.