© EYEWIRE

Dysphonic Speech Reconstruction

Construction of a Novel System for Effective and Efficient Communication

BY H. IREM TURKMEN AND M. ELIF KARSLIGIL

Verbal communication is one of the most influential and effective way of social communication. Voiced sounds are produced when the vocal cords vibrate; thus, the flow of air from the lungs to the vocal tract interrupts, and quasi-periodic pulses of air are produced during excitation. Dysphony is a functional disorder of larynx as a result of pathologic vibration in vocal cords. Chronic dysphonia occurs in the presence of organic lesions (such as polyp, nodule, and Reinke's edema) in the vocal cords, lethal larynx diseases, throat cancer, neurological disorders, and chronic irritation due to smoking. The breathed air, which is send to the trachea to produce voice, could make the vocal cords to vibrate barely or not at all because of the pathological formation in the vocal cords of the patients. As a result, voice comes out as a low whisper and more cracked than usual.

After total laryngectomy, there are three well-established methods to fix the voice. The first one is alaryngeal speaking. Through this method, the patient can speak by using an electrolarynx. The second method is training the patient and helping him speak in esophageal speech. The third method is speaking by using tracheoesophageal voice prostheses. Even though the speech is not as qualified as the previous, patients can speak through artificial larynx, voice prosthesis, or esophageal speech. However, these methods cannot be applied to the patients with apoplectic chordae vocalis, organic lesions of vocal cords, or who suffer from dysphony due to a partial laryngectomy in which some parts of the larynx and vocal cords are removed. Solutions such as voice therapies and/or operations to help patients to speak again may not work at all.

Several systems that analyze and enhance the characteristics of the esophageal speech and speaking using electrolarynx have been designed so far [1]–[5]. However, there is no reported research in the literature that produces synthetic voice digitally based on the patients' voice in cases where the patients were treated with partial laryngectomy or had completely lost speech as a result of organic lesions on the vocal cord or of vocal-cord paralysis.

In this article, we present a novel system that delivers synthetic speech with a quality close to natural by reconstructing dysphonic speech. We believe that it will be an important improvement in the social patients for effective and efficient communication.

Digital Object Identifier 10.1109/MEMB.2009.935725

Acoustic Characteristics of Dysphonic Speech

Chronical dysphonia mainly occurs because of the malfunctioning of the vocal cords. Voice formed this way demonstrates whisperlike characteristics. Dysphonic speech differs from normally phonated speech in terms of voicing, pitch, and formant structure. Spectrograms of normal and dysphonic speech for the Turkish word "Çalışma" (IPA codes of character ç = t$\int$ and ş = $\int$) are given in Figure 1.

Figure 1 clearly shows that, contrary to the voiced phonemes of normal speech, there is no perceivable pitch period or voicing observed in the voiced phonemes of dysphonic speech. In addition to this, voiced phonemes of dysphonic speech differ from the voiced phonemes of normal speech in terms of formant distortion. Bandwidths of dysphonic phonemes are larger, and their formant frequencies are greater. However, in unvoiced phonemes of dysphonic speech, there is no significant formant distortion observed [5]. Differences between dysphonic speech and normal speech are summarized in Table 1 in terms of pitch, voicing, and formant distortion characteristics.

According to Table 1, it was determined that no modification should be done for the unvoiced phonemes of a dysphonic speech.

Data Collection

The voices of dysphonic patients come out as whispers because their vocal cords cannot function properly. On the other hand, evaluating both the dysphonic voice and its original form before the disorder is essential to choose the appropriate method for normal speech reconstruction. Since accessing a dysphonic patients' original voice recordings is rather difficult, normal voices and whispers of healthy speakers were used to choose the proper method. For this purpose, a database consisting of recordings of normal voices and whispers of 30 men and 20 women speakers aged 25–50 was established.

There is no public database of dysphonic speech in literature; so, a dysphonic speech database containing 22 patients' speech recordings was created to appraise the success of the

Dysphony is a functional disorder of larynx as a result of pathologic vibration in vocal cords.

system. The recordings in the created database are of eight men and five women with apoplectic chordae vocalis, who were aged 35–55, and of six men and three women having partial laryngectomy, who were aged 45–65. The voice recordings were done in a silent room with a microphone that was placed 10-cm away from each patient's mouth. The sampling frequency was set at 8 kHz. The patients were made to read sentences that contained a total of 80 different Turkish words with normally distributed phonemes.

Dysphonic Speech-Enhancement System

Taking into consideration the acoustic differences of dysphonic and normal speech, a mixed excitation linear predictive coding (MELP) [6] based system was designed to enhance the dysphonic speech. Figure 2 illustrates the block diagram of the system.

First, speech of a dysphonic patient is recorded, and the recorded speech is analyzed by using MELP. Second, each phoneme is classified as voiced or unvoiced. To this end, mel-frequency cepstral coefficients are used to extract the features of voiced and unvoiced phonemes, and principle component analysis [7] is applied to reduce the dimensions of the features. Then, voiced and unvoiced phoneme groups are classified by using support vector machine [8].

Next, pitch, formant, and voicing parameters of voiced phonemes are modified. Synthetic pitch production is done by using the relation between pitch and gain values [9], [10].

Line spectrum frequency-based formant structure modification is applied to obtain narrow bandwidths and altered frequencies [11]. Line spectrum pair trajectories are smoothed by median filter during the vowels without destroying the rapidly varying spectral content of the phonemes [9].

Pulse excitation is produced for the voiced phonemes at a frequency band of 0–5 kHz. For this reason, the lower four frequency bands (0–3 kHz) are fixed as voiced and upper band (3–4 kHz) are fixed as unvoiced [9].

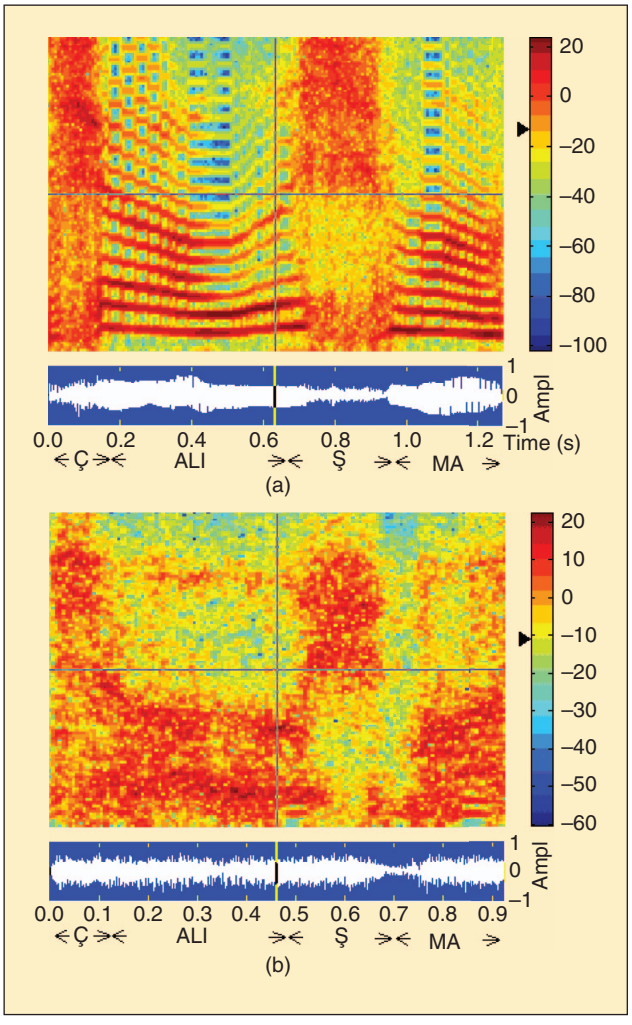

Fig. 1. Spectrogram of (a) normal speech and (b) dysphonic speech for ''Çalışma.''

Results and Observations

Log spectral distances were used to test the spectral differences between the normal and synthetic speech. The average spectral enhancement obtained was calculated as 25%. Figure 3(c) and (d) shows two different synthetic speech spectrograms for the Turkish word ''Taş'' spoken by a patient with partial laryngectomy. In Figure 3(d), all the phonemes, voiced and unvoiced, were modified, whereas in Figure 3(c), only the voiced ones.

As can be seen in Figure 3(c), with the help of the modifications, synthetic speech was given the attribute of periodicity

Table 1. Acoustic differences of dysphonic and normal speech.

	Pitch	Voicing	Formant Distortion
Dysphonic speech			
Unvoiced phonemes	×	×	×
Voiced phonemes	×	×	√
Normal speech			
Unvoiced phonemes	×	×	×
Voiced phonemes	√	√	×

√: Presence of the related feature; ×: absence of the related feature.

similar to the ones of normal speech [Figure 3(a)]. The results show that preservation of the acoustic features of unvoiced phonemes produce a synthetic speech that is closer to the normally phonated one.

To determine the preferentiality of the synthesized speech versus dysphonic speech, subjective listening tests were conducted. Fifteen listeners were requested to listen to the samples and then rate them within a scale of 1–5 (1 being the worst and 5 the best) according to timber, intelligibility, and naturalness. The calculated average points are given in Table 2.

As can be seen in Table 2, synthetic speech produced by the use of either approaches is recognized better than dysphonic speech, especially in terms of timber and naturalness.

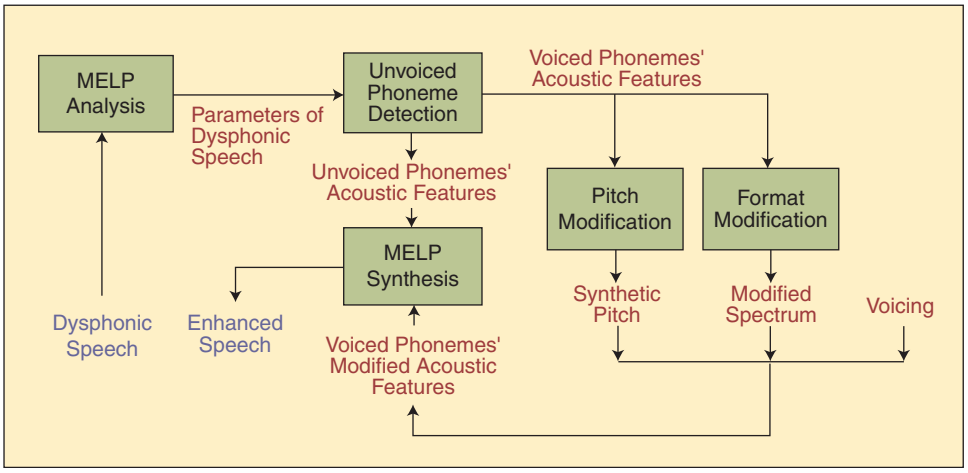

Fig. 2. Block diagram of the proposed system.

Future Work

There are currently no commercially available devices to use normal speech reconstruction from dysphonic speech. Our normal speech reconstruction software can be implemented in

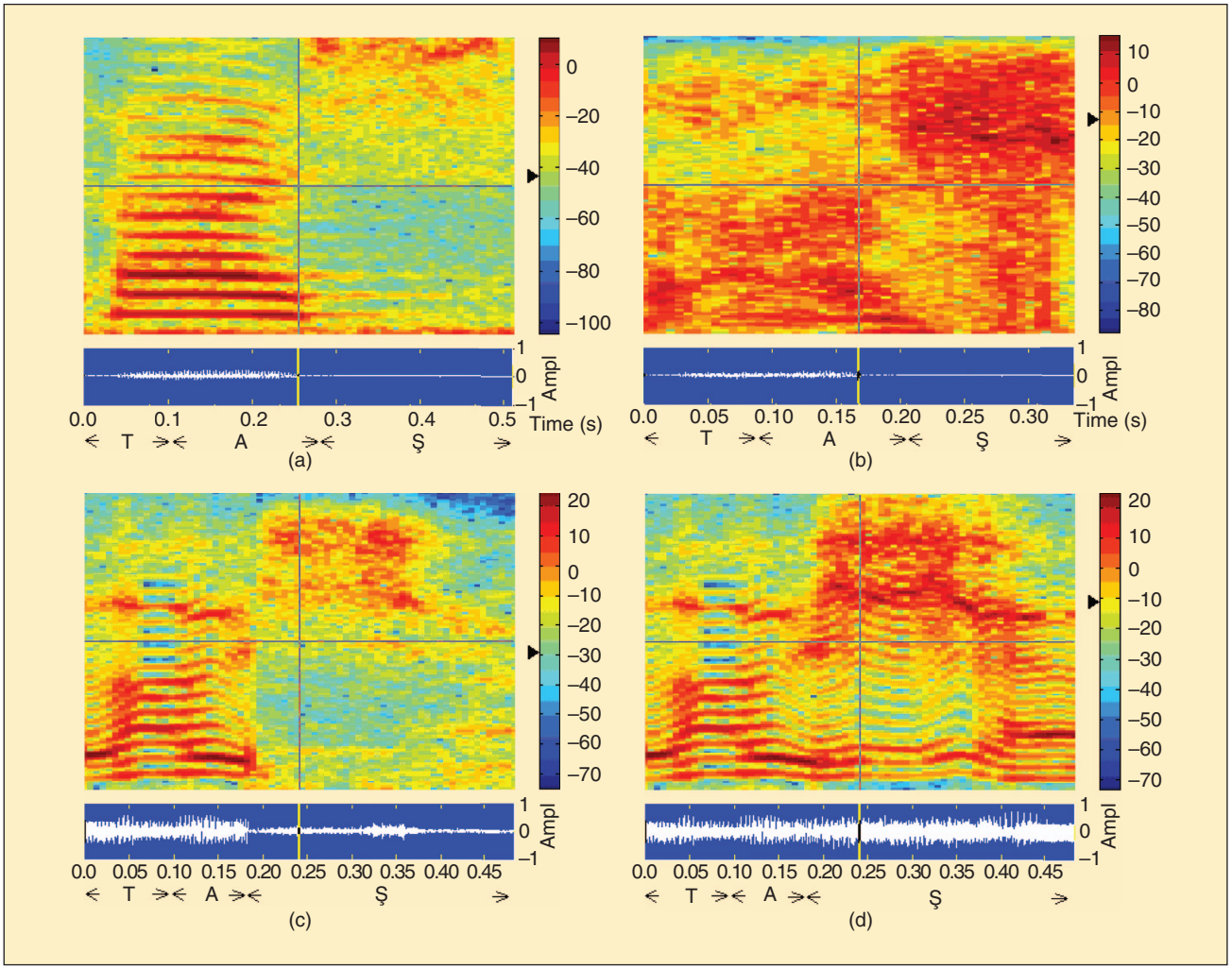

Fig. 3. Spectrograms of Turkish word TAŞ: (a) normal speech, (b) dysphonic speech, (c) synthetic speech produced by modification of voiced phonemes, and (d) synthetic speech produced by modification of each phoneme.

Table 2. Comparison of dysphonic speech with the synthetic speech.

	Timber	Intelligibility	Naturalness
Dysphonic speech	0.3	2.3	1.1
Synthetic speech produced by modification of each phoneme	2.8	2.4	3.0
Synthetic speech produced by modification of voiced phonemes	2.9	2.8	3.1

an embedded system of the size of a mobile MP3/4 player or a cellular phone, which patients could carry with them all the time. One can also think of implementing it directly on a commercially available cellular phone as a software add on. Considering the capabilities of cellular phones like recording and playback of streaming audio and video, this processing power can be used for normal speech reconstruction. It is obvious that the presented algorithm should be tailored and/or optimized for the aforementioned resource constrained-embedded devices, which, we believe, is not of great concern. In addition, cellular phones are already equipped with the right voice recording and playback hardware. Therefore, the normal speech reconstruction software installed on a cellular phone can intercept all outgoing speech data during an active call, reconstruct the dysphonic speech, and then have it sent to the calling party. Depending on the processing power and/or memory capacity of the cellular phone, it may be necessary for the speaker to give brief pauses every one or two sentences for the software to do the reconstruction.

As stated earlier, the normal speech reconstruction software could also be implemented on a dedicated embedded hardware as a standalone mobile device. This standalone device could have a built-in speaker for close-range conversations and/or both analog and digital sound outputs to be tied to a sound system for using it when the speaker needs to address bigger crowds. In this scenario, the mobile device could be equipped with necessary processing power and memory capacity to enable a streaming speech reconstruction as the speaker speaks freely.

Especially, both the mobile communications and entertainment industries are the driving force for the hardware manufacturers to produce more powerful CPUs and higher density memories with lesser power consumption available at lower prices. Therefore, we believe that it is justified to think that the normal speech reconstruction system could be made available to the patients at very reasonable prices both as a software-only solution available in the form of a software add on, e.g., cellular phones, as a dedicated mobile device.

Conclusions

In this article, we present a MELP-based real-time system that produces synthetic speech for patients with chronic dysphonia while preserving their continual structure of speech.

Dysphonic patients, even though strained, could express themselves within face-to-face communications. However, when speaking in a group or using communication devices such as mobile phones, they face serious difficulties in vocal communication since it becomes harder to understand their speech. Our system will make the speech of such a patient much more understandable by reconstructing the patients' dysphonic speech in real time. Our system can easily be fit into an embedded system of the size and shape of a mobile phone, so that it could be carried all the time by the patients.

Acknowledgment

We express our appreciation to Istanbul University Cerrahpasa Medical Faculty—Ear Nose Throat and Head & Neck Surgery Department for their support in this work.

H. Irem Turkmen received her B.Sc. and M.Sc. degrees in computer engineering from Yildiz Technical University, Istanbul, Turkey, in 2005 and 2008, respectively. She is a Ph.D. student at the Computer Engineering Department of Yildiz Technical University. Her research interests include pattern recognition, speech processing, and machine learning.

M. Elif Karsligil received her B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Yildiz Technical University, Istanbul, Turkey, in 1988, 1990, and 1998, respectively. She is currently an assistant professor at the Computer Engineering Department of Yildiz Technical University. From October 2001 to November 2002, she worked as senior researcher at NTT Communication Science Laboratories, Kyoto, Japan. Her research interests include machine learning, pattern recognition, speech processing, and digital video processing.

Address for Correspondence: H. Irem Turkmen, Yildiz Technical University, Computer Engineering Department, 34349 Yildiz, Istanbul, Turkey. E-mail: irem@ce.yildiz.edu.tr.

References

[1] G. Aguilar and M. Nakano-Miyatake, "Alaryngeal speech enhancement using pattern recognition techniques," *IEICE Trans. Inform. Syst.*, vol. 88, no. 7, pp. 1618–1622, 2005.

[2] N. Bi and Y. Qi, "Speech conversion and its application to alaryngeal speech enhancement," in *Proc. 3rd Int. Conf. Signal Processing (ICSP'96)*, 1997, pp. 1586–1589.

[3] A. Pozo and S. Young, "Continuous tracheoesophageal speech repair," in *Proc. European Signal Processing Conf. (ICSP'96)*, Florence, Italy, Sept. 4–8, 2006.

[4] Y. Qi, B. Weinberg, and N. Bi, "Enhancement of female esophageal and tracheoesophageal speech," *J. Acoust. Soc. Amer.*, vol. 98, no. 5, pp. 2461–2465, 1995.

[5] H. Sawada, N. Takeuchi, and A. Hisada, "A real-time clarification filter of a dysphonic speech and its evaluation by listening experiments," in *Proc. Int. Conf. Disability, Virtual Reality and Associated Technologies (ICDVRAT'04)*, 2004, pp. 239–246.

[6] A. V. McCree and T. P. Barnwell, III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

[7] V. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Predicting F0 and voicing from NAM-captured whispered speech," in *Proc. 4th Conf. Speech Prosody*, Campinas, Brazil, May 2008, pp. 107–110.

[8] V. N. Vapnik, "An overview of statistical learning theory," *IEEE. Trans. Neural Networks*, vol. 10, no. 5, pp. 998–999, Sept. 1999.

[9] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, no. 7, pp. 515–520, 2002.

[10] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels- relationship with formant frequencies: A preliminary study," *J. Voice*, vol. 10, no. 2, pp. 155–158, June 1996.

[11] I. V. McLoughlin and R. J. Chance, "LSP-based speech modification for intelligibility enhancement," in *Proc. 13th Int. Conf. DSP*, vol. 2, pp. 591–594.