Reconstruction of Dysphonic Speech by MELP

H. Irem Türkmen and M. Elif Karsligil

Yildiz Technical University, Computer Engineering Department,
34349 Yildiz, Istanbul, Turkey
{irem,elif}@ce.yildiz.edu.tr

Abstract. The chronical dysphony is the result of neural, structural or pathological effects on the vocal cords or larynx and it causes undesirable changes in the quality of speech. This paper presents a Mixed Excitation Linear Prediction (MELP) based system that reconstructs normally phonated speech from dysphonic speech, while preserving the individuality of the patient. The proposed system can be used as speech prosthesis for the patients who have lost the ability to produce voice. To reconstruct normally phonated speech from dysphonic speech, pitch generation using the perceived pitch relationship with formant frequencies, formant and voicing modification steps were performed for phonemes. The principle novelty of this study is to modify voiced phonemes' acoustic features while preserving unvoiced ones. Therefore voiced-unvoiced detection is performed for each phoneme.

The proposed system is composed of three main parts. In the analysis phase the acoustic differences observed between normal and dysphonic speech are determined. Acoustic parameters of the dysphonic speech's voiced phonemes are modified in order to obtain a synthetic speech that is closer to normal speech. Finally, enhanced speech is synthesized by MELP.

Keywords: Dysphonic speech enhancement, MELP, Formant modification, Pitch and voicing generation.

1 Introduction

Verbal communication is one of the most influential and effective way of social communication. While producing voice, airflow from the lungs to the vocal tract is interrupted by the vibration of vocal cords and quasi-periodic pulses of air are produced as the excitation.

The chronical dysphony occurs in the presence of organic lesions, vocal cord paralysis, larynx cancer and results in the loss of ability to speak. Surgery for laryngeal cancer results in the removal of the larynx including vocal cords. During laryngectomy, surgeon perforates a hole in patient's neck called stoma that the patient can breathe through. After surgery, oesophageal, electrolarynx and the tracheoesophageal (TE) speech are the ways to speak. However these techniques have disadvantages. The major drawback with esophageal speech is that the sounds are rough and often limited to relatively short segments of speech. The electrolarynx has a very mechanical tone that does not sound natural and good hand control is required to use the electrolarynx. TE voice prosthesis must be removed and cleaned periodically because

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 767–774, 2008.

infection risk exists [1]. The main purpose of this research is to developing a dysphonic speech enhancement system that can be used as speech prosthesis for the patients who have lost the ability to produce voice.

Several researches which analyze and enhance the characteristics of the oesophageal and electrolarynx speech have been reported so far [2-6]. Morris and Clements [7] proposed a system that modifies formant structure and determines pitch and voicing to reconstruct speech from whisper by using MELP.

In the proposed system, Turkish speech samples were recorded from native Turkish speakers who have had their larynx removed or have paralyzed vocal cords. MELP is used for synthesizing enhanced speech. Pitch relationship with formant frequencies is used in order to produce pitch for dysphonic voice. The system is composed of three major parts: Analysis of the dysphonic speech, modification of the acoustic parameters of the dysphonic speech in order to obtain synthetic speech which is closer to normal speech and finally, synthesizing the enhanced speech using the modified parameters. Modification was not applied to unvoiced phonemes, since there is no significant distortion observed in dysphonic speech for unvoiced phonemes. Figure 1 shows the block diagram of the proposed system.

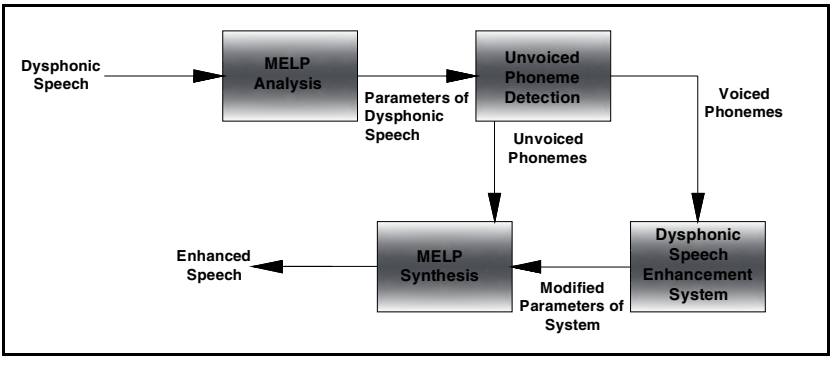


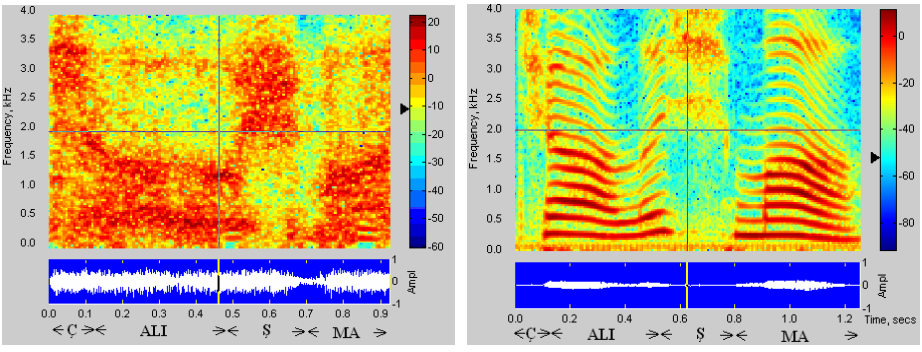
Fig. 1. Block Diagram of Proposed Speech Reconstruction System

2 Acoustic Differences between Dysphonic and Normally Phonated Speech

Dysphonic speech differs from normally phonated speech in terms of voicing, pitch and formant structure. There is no perceived pitch period in dysphonic speech and the voice is definitely noisy. Two spectrograms for the Turkish word “calisma” (IPA Code of character c=CH, s=SH [8]) are given in Figure 2. The spectrogram in Figure 2a belongs to a patient with paralyzed vocal cords whereas Figure 2b shows the spectrogram of the normally phonation of the same word.

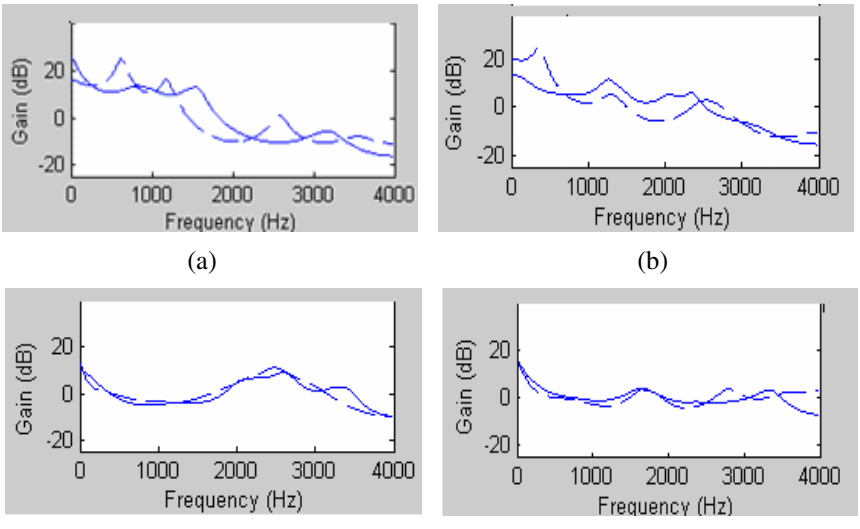
Several studies demonstrate that the formant locations and bandwidths of dysphonic speech differ from normally phonated speech [4]. LPC spectra of dysphonic (solid line) and normally phonated (dashed) phoneme samples are shown in Figure 3.

As it can be seen in Figure 3a-3b, a formant structure distortion is observed in voiced phonemes, while there is no significant distortion observed in unvoiced ones



(a) (b)

Fig. 2. Spectrogram of (a) Dysphonic Speech (b) Normal Speech



(a) (b)

(c) (d)

Fig. 3. LPC spectra of dysphonic and normal voice for the phonemes (a) /AA/ as in dArk (b) /r/ as in Rate (c) /k/ as in Coat (d) /s/ as in Sue

(Figure 3c-3d). Moreover, it is observed that, voiced frequency bands of the unvoiced phonemes, which are pronounced by a dysphonic speaker, and normal words are not different contrary to the voiced frequency bands of voiced phonemes. Also unvoiced phonemes have no perceived pitch when they are pronounced by a normal speaker.

3 Dysphonic Speech Enhancement System

As suggested in part 2, no perceived pitch, and excitation exist in dysphonic speech. Also formant structure distortion is observed. In order to enhance the dysphonic

speech, voicing decision, pitch estimation, gain and formant structure modification should be applied. On the other hand, applying the same procedure to unvoiced phonemes decreases intelligibility. As a novel approach, the proposed system modifies the acoustic parameters of phonemes except unvoiced phonemes to increase the synthetic speech quality.

3.1 Detection of Unvoiced Phonemes

The need for classifying a given speech segment as voiced or unvoiced arises in many speech analysis systems. Pitch analysis, autocorrelation function and zero crossing rate are usually the methods used to make voiced-unvoiced decision [9]. However, since there is no perceived pitch observed in dysphonic speech, it is hard to make voiced-unvoiced decision using pitch analysis. In addition to this, autocorrelation coefficients and zero crossing rates are not distinctive features for voiced-unvoiced classification.

In the proposed system, speaker dependent classification of voiced and unvoiced phonemes was made by using line spectrum frequencies. We manually constructed two classes of phonemes with respect to their articulation. First class contains unvoiced phonemes, and the second one contains voiced phonemes. Train set consists of the average line spectrum frequencies of voiced and unvoiced dysphonic phonemes. K-Nearest Neighborhood was applied by cross validation technique for the detection of unvoiced phonemes. The classification accuracy for phoneme groups for k =3 is given in Table 1. Analysis of the classification errors showed that about 48 percent of the errors occurred when classifying voiced consonants z, r, j and g whereas about 2 percent of errors were observed for y, v, m, n, l, d and SH. Moreover we observed that the system frequently misclassified unvoiced fricative phonemes HH and p. In the proposed system acoustic parameters of voiced phonemes were modified while acoustic parameters of unvoiced phonemes' were preserved.

Table 1. Classification accuracy of phoneme groups

	Vowels	Voiced Consonants	Unvoiced Consonants
Classified as Unvoiced Consonant	5,12%	17,23%	74,38%
Classified as Voiced Consonant or Vowel	94,88%	82,77%	25,62%

3.2 Voicing Decision

The proposed method fixes the lower four frequency bands (0 – 3 kHz) as voiced, while fixing the upper band (3 – 4 kHz) as unvoiced [7].

3.3 Pitch Estimation

Dysphonic speech has no perceived pitch. The synthetic speech should be natural. In order to accomplish this goal, a pitch estimation process was applied to voiced speech segments. By using the observed correlation between intensity and perceived pitch, the pitch parameter was estimated by the following equation with $pitch_{new}^{n}$, estimated

pitch, $gain^{n}$, gain of the frame number n, $gain_{average}$, average gain of dysphonic speech segment, $pitch_{reference}$, reference pitch [7]. While $pitch_{reference}$ is used to adjust the tone of the synthetic speech, β is used to adjust the dynamic range of the pitch period.

$$pitch^{n}_{new} = ((gain^{n} - gain_{average}) * \beta) + pitch_{reference} \quad (1)$$

In the proposed system, $pitch_{reference}$, is calculated automatically. Since it is too hard to obtain the normal voice of the dysphonic speaker and like dysphonic speech, whispered speech has no perceived pitch period, second formant frequency of whispered /AA/ phoneme is used to calculate the most appropriate pitch for the dysphonic speaker.

Several studies point out a relationship between pitch and formant frequencies [10, 11]. To formulate the relationship, formant frequencies of /AA/ phoneme, which belong to different speakers, were studied.

Spectra of normally phonated /AA/ phoneme that are voiced by four speakers who have various voice tones are shown in Figure 4. The pitch periods of the speakers are calculated as 20, 36, 52 and 89 by using normalized autocorrelation function. As seen in Figure 4, while pitch period increases, second formant frequency decreases.

Spectra of the whispered versions of the same phoneme are shown in Figure 5.

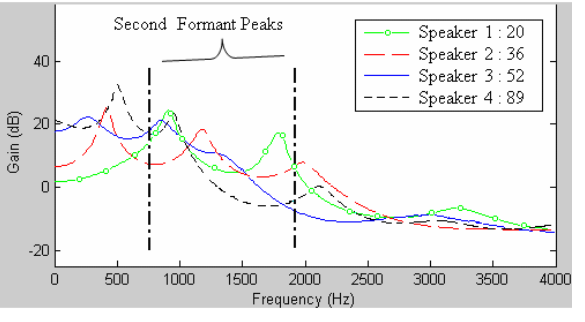


Fig. 4. Spectra of normally phonated /AA/ phoneme that are voiced by four speakers

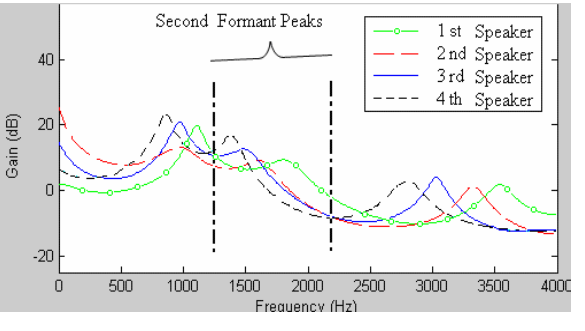


Fig. 5. Spectra of whispered /AA/ phoneme that are voiced by four speakers

As it is evident from Figure 5, second formant frequency of the whispered phoneme /AA/ voiced by speaker 1 with lowest pitch period is highest. Pitch and formant frequency are inversely proportional.

Reference pitch $pitch_{reference}$ can be calculated by using the following equations where $f^{highest}$ is the second formant frequency of the speaker who has the highest pitch and $p^{highest}$ is the pitch of that speaker and f^{lowest} is the second formant frequency of the speaker who has the lowest pitch and p^{lowest} is the pitch of that speaker.

$$a = (p^{highest} - p^{lowest}) / (f^{lowest} - f^{highest}) \tag{2}$$

$$p_{referans} = (f^{lowest} - f_2) * a + p^{lowest} \tag{3}$$

In the proposed system, $pitch_{reference}$ is calculated by using the pitch and second formant frequency values of the speakers in train set who have the highest and the lowest pitch. Hence, $f^{highest}$, $p^{highest}$, f^{lowest} and p^{lowest} were set to 897, 89, 1788 and 20 respectively.

3.4 Formant Structure Modification

In the proposed system, LSF based formant structure modification is applied to obtain narrow bandwidths and altered frequencies [12]. LSP trajectories are smoothed by median filter during the vowels without destroying the rapidly varying spectral content of the phonemes, [7].

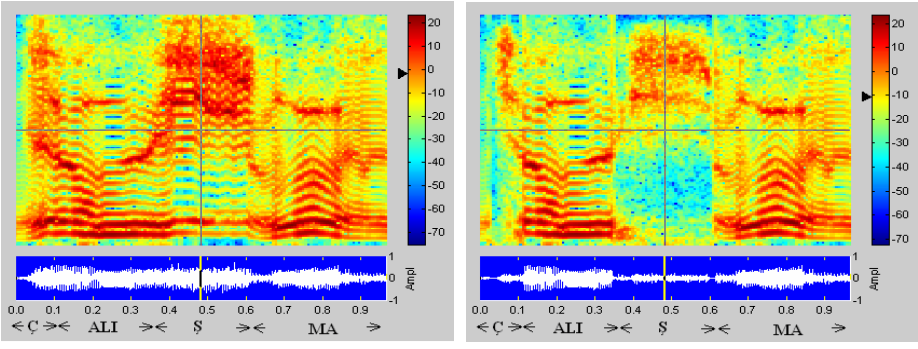
4 Experimental Results

In this study, 50 triphone-balanced sentences were recorded from 5 male and 2 female dysphonic Turkish native speakers.

Preserving the acoustic features of unvoiced phonemes increases the intelligibility of the synthetic speech. Figure 6a shows the spectrogram of the synthetic speech for the dysphonic word “calisma” (Figure 2a) produced by the modification of every phoneme, whereas Figure 6b shows the spectrogram for the same word produced by the modification of only voiced phonemes.

As it is evident from Figure 6, preservation of the acoustic features of unvoiced phonemes results in synthetic speech that is closer to normally phonated one.

In order to test the spectral differences between normal and synthetic speech, log spectral distances were used. Acquired average spectral enhancement is calculated as %25. Because the spectral difference is only one part of the conversion, subjective testing was also applied to evaluate how well we can synthesize normal speech from dysphonic speech. 5 listeners were asked to vote the synthetic speech in terms of the intelligibility and similarity to normal speech as 5 is best.



(a) (b)

Fig. 6. Spectrogram of synthetic speech for word "calisma" (a) produced by modification of each phoneme (b) unvoiced phonemes acoustic features preserved

Table 2. Subjective listening test results

	intelligibility	normal speech similarity
Original Dysphonic Speech	2. 1	1. 1
Enhanced Speech	2. 7	2. 5

5 Conclusion

This paper presents a MELP based system that enhances dysphonic speech. To reconstruct normal speech from dysphonic speech, pitch generation, formant and voicing modification steps were applied to only voiced phonemes, leaving the unvoiced phonemes unmodified.

Subjective listener tests indicate the distinct similarity between synthetic speech and normally phonated speech. Adjusting the modification of the formants according to the phoneme structure and computing more natural pitch contours would increase the success rate.

Our proposed system could be used to improve the life quality of dysphonic patients in every day situations like telecommunication applications.

Acknowledgements

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

References

1. Eastern Virginia Medical School, `http://www.evmsent.org`
2. Aguilar, G., Nakano-Miyatake, M.: Alaryngeal Speech Enhancement Using Pattern Recognition Techniques. IEICE - Transactions on Information and Systems E88-D(7), 1618–1622 (2005)

3. Bi, N., Qi, Y.: Speech conversion and its application to alaryngeal speech enhancement. In: Proc. ICSP 1996, pp. 1586–1589 (1997)
4. Sawada, H., Takeuchi, N., Hisada, A.: A Real-time Clarification Filter of a Dysphonic Speech and Its Evaluation by Listening Experiments. In: International Conference on Disability, Virtual Reality and Associated Technologies (ICDVRAT 2004), pp. 239–246 (2004)
5. Pozo, A., Young, S.: Continuous Tracheoesophageal Speech Repair. In: EUSIPCO (2006)
6. Qi, Y., Weinberg, B., Bi, N.: Enhancement of female esophageal and tracheoesophageal speech. Journal of the Acoustical Society of America 98, 2461–2465 (1995)
7. Morris, R.W., Clements, M.A.: Reconstruction of speech from whispers. Medical Engineering and Physics 24(7), 515–520 (2002)
8. The International Phonetic Association, `http://www.arts.gla.ac.uk/ipa/fullchart.html`
9. Atal, B.S., Rabiner, L.R.: A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing assp-24(3) (June 1976)
10. Thomas, I.B.: Perceived pitch of whispered vowels. J. Acoust. Soc. Am. 46(2), 468 (1969)
11. Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H.: Perceived pitch of whispered vowels- relationship with formant frequencies: A preliminary study. Journal of Voice, 155–158 (1996)
12. McLoughlin, I.V., Chance, R.J.: LSP-based speech modification for intelligibility enhancement. In: Proceedings 13th International Conference on DSP, vol. 2, pp. 591–594 (1997)