ENHANCEMENT OF ESOPHAGEAL SPEECH USING FORMANT SYNTHESIS

Kenji Matsui and Noriyo Hara

Central Research Laboratories, Matsushita Electric Ind. Co., Ltd.
3-4 Hikari-dai, Seika, Soraku
Kyoto, 619-0237, JAPAN
matsui@crl.mei.co.jp and haran@crl.mei.co.jp

ABSTRACT

The feasibility of using the formant analysis-synthesis approach to replace the voicing sources of esophageal speech was explored. The voicing sources were generated by using inverse-filtered signals extracted from normal speakers. Pitch extraction was tested with various pitch extraction methods, then simple auto-correlation method was chosen. Special hardware unit was designed to perform the analysis-synthesis process in real-time. Results of a subjective test showed that the synthesized speech was significantly improved.

1. INTRODUCTION

Person who have had laryngectomies have several options for the restoration of speech, none completely satisfactory. The artificial larynx, typically a hand-held device which introduces a source vibration into the vocal tract by vibrating the external walls, is the easiest for patients to master, but does not produce airflow, so that the intelligibility of consonants is diminished. Tracheo-esophageal speech, which utilizes a prosthesis to divert outgoing lung air into the esophagus, bringing about a vibration of the esophageal superior sphincter, provides airflow for consonants and permits utterances of normal duration. However, it requires a surgically produced connection between the esophagus and the trachea, and is not suitable for some patients. Esophageal speech, which requires speakers to insufflate, or inject air into the esophagus, limits the pitch range and intensity. Both esophageal speech and tracheo-esophageal speech are characterized by low average pitch frequency, large cycle-to-cycle perturbations in pitch frequencies, and low average intensity. The formant patterns of esophageal speech, however, are found to be similar to those of normal speakers, except for overall elevations of formant frequencies attributed to the shortening of the vocal tract as a result of the surgical method. A system that converts esophageal speech into normal speech could be useful to enhance communication for alaryngeal talkers. It is well documented that LPC analysis is an effective tool for estimating the parameters of the linear system for normal speech production. To enhance the quality of esophageal speech, Qi attempted replacing the voicing source of esophageal speech using a LPC method[1],[2],[3],[4].

This study was undertaken to explore the feasibility of using the LPC method to replace the voicing source of esophageal speech. Also, an enhancement device for esophageal speech running in real-time was designed and evaluated with the LPC method to figure out the actual speech conversion performance under the real environment. The specific goals were:(1)to determine the algorithm and control parameters for analysis-synthesis of esophageal speech. (2)to determine if the synthesized speech has high intelligibility; and (3) to determine if the hardware unit is able to make esophageal speakers easier to communicate.[5],[6],[8]-[13]

2. SPEECH ANALYSIS-SYNTHESIS

Our method for analyzing esophageal speech is based on relatively straightforward formant analysis synthesis method with parameter smoothing techniques. The basic scheme is shown in Figure 1. The input speech signal is digitized by sampling at 10kHz. The input signal is divided into two channels at 2.5kHz with 4^{th} order low pass and high pass filters.. The only low frequency channel is used for the synthesis processing. The high frequency channel is mixed with the synthesis filter output at the final stage. The signal is pre-emphasized and is used for processing. Every 10msec 10^{th} order LPC analysis and simple auto-correlation-based pitch analysis are performed on a 30ms window of speech samples. The fundamental frequency range is pre-selected to avoid the unstable pitch extraction result. The first three formants are extracted using Bairstaw method. The only power level is used to discriminate voiced and unvoiced speech frames considering the unstable speech input. The unvoiced speech frames are not processed. Only voicing

frames are analyzed and synthesized with pre-selected inverse-filtered voicing source. The new excitation sources are generated from one-pitch-period glottal waveform samples extracted from normal talkers, and the fundamental frequency is able to increase or decrease by preset condition[7]. The analyzed parameters (pitch, power, formant frequencies, band width) are filtered by 3 point median smoothing. Fourth formant and fifth formant are usually set to constant values based on the analysis result of the users. The output from the synthesis filters is merged with the output from the higher frequency channels.

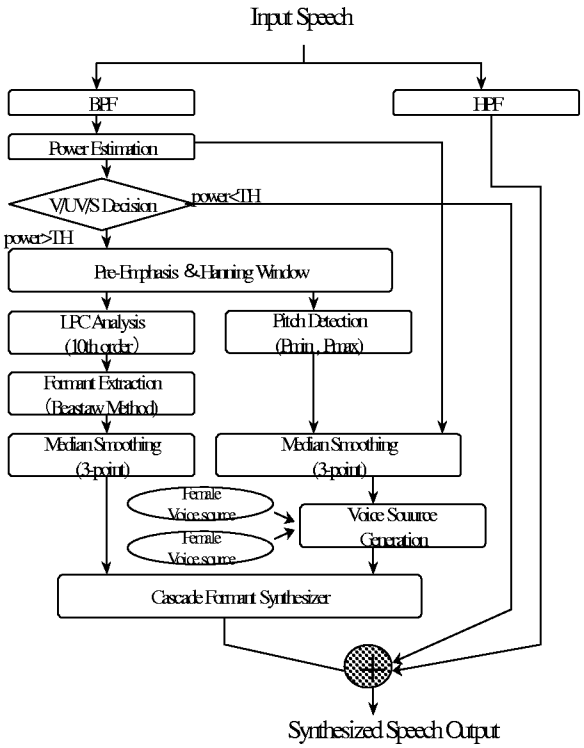

Figure 1. Block diagram of the formant analysis-synthesis method.

3. DSP UNIT

A block diagram of the hardware architecture is shown in Figure 2. The PC based speech analysis tools have been developed to analyze all of the necessary parameters, (ie. average power, pitch range, V/UV power threshold) in advance. We have introduced TMS320C32 digital signal processor to realize all of the analysis synthesis process in real time. The DSP hardware unit consists of a small board with a 60MHz C32, a 32 bit floating point DSP, a

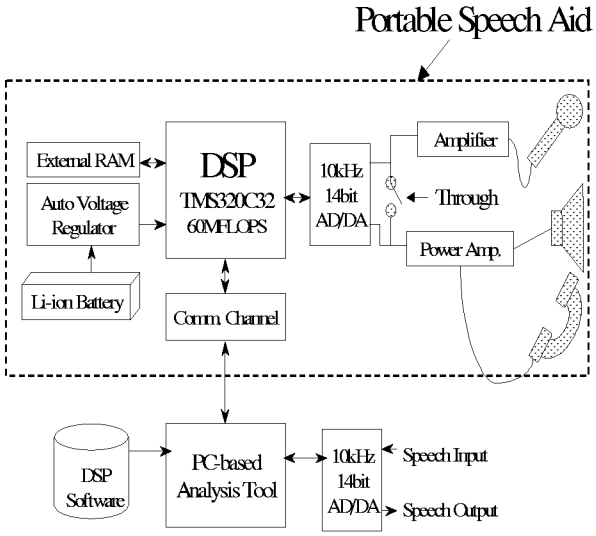

Figure 2. Block diagram of the protable speech aid hardware architecture.

fourteen bit A/D and D/A with 10kHz sampling, an parallel host interface, 16Mbit SRAM 20ns access time, analog circuits for a microphone and laud speaker, auto-voltage regulator, Li-ion battery. With the combination of the PC-based speech analysis tools and the DSP hardware unit, a new user can first adjust all of the necessary parameters, then download the adjusted DSP software into the hardware unit. The overall signal processing delay is unavoidable. Because of the analysis frame length (30ms) and the three point median smoothing, the system requires 40ms delay.

4. EXPERIMENTAL RESULTS

Subjective evaluation tests(rating scale method) have been made with 5 male and 2 female well-trained esophageal speakers, and 3 speech therapists. Each speaker first speaks “aoi o ueru” twice to adjust all of the necessary parameters.(Figure 5) After the pre-setting of the hardware unit, the speaker reads the test materials without using the enhancement device first, then speaks same materials with the device. The test materials consist of seven sentences considering phoneme categories.

(1) aoio ueru
(2) anohitowa bunkajinto yobarerunoga fusawashii
(3) shichigatsukara hanshindenshade tsukinshiteimasu
(4) ginkomo gakkomo aruiteikeru kyorini arimasu
(5) kinkoga toreteirunode kakkoga yoi
(6) shizyu gamuo kamunoga shukan ni natteiru
(7) hanao ottari anao hottari sanzanna meniatta.

The speech therapists rate the quality of the unprocessed speech stimuli and the processed ones using multidimensional scaling. The t-test result of the subjective evaluation indicated that the synthesized speech obtained better scores except one female speaker who has almost normal speech quality. (Table 1., Figure 3.)

5. DISCUSSION

The results of this study indicate that the speech quality of esophageal speaker could be improved by LPC based formant analysis and synthesis process. The ability to improve the voice quality with proposed method implies that the glottal source replacement may be adequate for the esophageal vowel reproduction. The quality improvements were observed in the eleven features out of nineteen ones. No significant improvement, however, was found on "Breathy" feature, although we had expected to obtain much better result. This was because one of the female speaker gave "breathy" impression when the fundamental frequency was increased. On the other hand, "choppy" ,"strain" and "duration" features, which are not related to the signal processing effect, obtained high scores. Even just amplifying the input speech may relieve the speaker's tension. The stoma noise was significantly reduced because the close-talking microphone characteristics effectively reduce ambient noise. Every speakers were asked to fill out a form to state their impressions. Some important topics are:

(1) The device may require some training period to feel comfortable.
(2) Increasing the original fundamental frequency gives better sound impression.
(3) The processed speech sounds like some other people's voice.

Also, the speech therapists were also asked to summarize their impression.

(1) Loudness and overall speech quality are improved.
(2) The device is effective especially for hoarse or breathy voice.
(3) Sometime the injection noise is amplified.
(4) The input and output mixed voice is heard due to the time delay.

In the case of female speech, increasing the original pitch gives better sound.

6. CONCLUSION

LPC based formant analysis synthesis method for esophageal speech was proposed, and the real-time DSP unit was designed. The device was tested in a practical environment by speech therapists. Results of the subjective evaluation indicated that the processed speech was preferred in most of the evaluation criteria. However, the esophageal speakers may be required to do some training before utilizing the device comfortably. The device also needs farther evaluation from both listener and speaker side under various practical conditions.

7. ACKNOWLEDGMENTS

The authors would like to express our gratitude to Prof. H.Hirose and Prof.. N.Kobayashi for their great help to make the esophageal speech database, the subjective evaluation and the useful discussion. We would like to thank the members of GINREIKAI for their cooperation with the recording and the subjective evaluation. We also would like to thank Mr. K.Ohira , Ashida Sound Co., Ltd. Mr. Kubota and Ms. M.Jin , TEAC Corporation that they are collaborating on this work with us.

This Work was performed as a part of the Research & Development of Medical & Welfare Apparatus supported by NEDO (New Energy and Industrial Technology of Development Organization) of Japan.

8. REFERENCES

[1] H.Javkin, M.Galler, and N.Niedielski, " Enhancement of esophageal speech by injection noise rejection", Proc. IEEE ICASSP pp1207-1210 (1997)

[2] N.Bi and Y.Qi, " Application of speech conversion to alaryngeal speech enhancement.", IEEE Trans. on speech and audio proc. Vol.5 pp.97-105 No.2 (1997. 3)

[3]Y.Qi, B.Weinberg, " Characteris of voicing source waveforms produced by esophageal and tracheoesophageal speakers. " J. Speech Hear. Res., Vol.38 pp.536-548 (1995.6)

[4]Y.Qi, " Replacing traceoesophageal voicing source using LPC synthesis", J. Acoust. Soc. Am. 88(3) pp1228-1235 (1990.9)

[5]K.Matsui and N.Hara, "Development and evaluation of speech aid system for esophageal speakers." Proc. of fall meeting of ASJ 2-P-24 (1998.9)

[6]N.Hara and K.Matsui, "Pitch detection performance for esophageal speech and Amplitude to pitch-contour Conversion Method", Proc. of fall meeting of ASJ 2-P-23 (1998.9)

[7]S.Peason, H.Javkin, K.Matsui, and T.Kamai, "Text-to-speech synthesis using a natural voice source." ICSLP90 6.8.1 pp.193-196 (1990)

[8]N.Hara and K.Matsui, " V/UV/N detection by using discriminant function for esophageal speech.", Proc. of spring meeting of ASJ 2-P-11 (1998.3)

[9]N.Hara and K..Matsui, " Adaptive gain control method for esophageal speech.", Proc. of fall meeting of ASJ 1-P-13 (1997.9)

[10]K.Matsui and N.Hara, " Real-time formant analysis-synthesis method for esophageal speakers. ", Proc. of fall meeting of ASJ 1-P-14 (1997.9)
[11]K.Matsui, E.Noguchi, and Y.kato, " Enhancement of esophageal speech using formant synthesis method.", Proc. of spring meeting of ASJ 1-Q-7 (1997.3)
[12]K.Matsui, E.Noguchi, and H.Javkin, " Enhancement of esophageal speech.", Proc. of fall meeting of ASJ 2-6-14 (1997.9)
[13]E.Noguchi and K.Matsui, " An evaluation of esophageal speech enhancement.", Proc. of fall meeting of ASJ 2-6-13 (1997.9)

Skill	Excellent					Intermediate	
Speaker	W	M I	S	A	H	N	MO
Sex	Male	Male	Male	Female	Female	Male	Male
LPC order	1 0	1 0	1 0	1 0	1 0	1 0	1 0
F 4	Fixed	Fixed	Fixed	Extracted	Extracted	Fixed	Fixed
F0 detection boundaries	60～150	60～150	60～150	60～130	60～130	60～130	60～120
Power Threshold	30dB	30	30	40	30	30	30
F0-shift	1	1	1	2	2	1	1
HPF enhance rate	1	1	1	1	1	1	1
t-*TEST*	$p < 0.01$	$p < 0.1$	$p < 0.01$	NSD	$p < 0.01$	$p < 0.01$	$p < 0.01$

Table 1. Preset parameter values of each esophageal speaker.

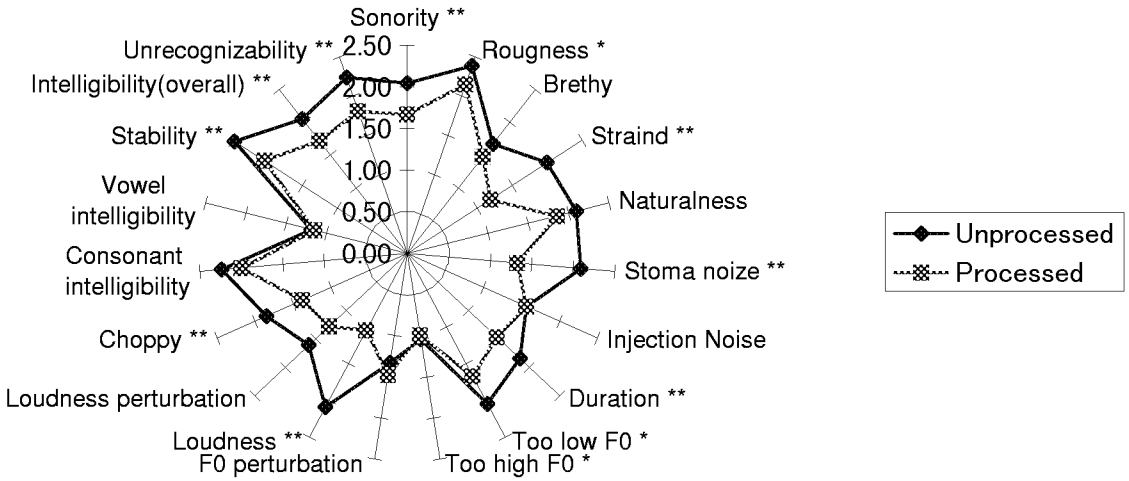

Figure 3. Subjective evaluation result(The average scores from 3 speech therapists)
Intelligibility 1 : Clear, 2 : Almost clear, 3 : Clear if known phrase, 4 : Almost unclear, 5 : Very unclear
Other items 1 : Normal, 2 : Slightly Annoying (light) , 3 : Annoying (middle) , 4 : Very Annoying (severe)