A New Feature Driven Cochlear Implant Speech Processing Strategy

Dashtseren Erdenebat, Kitazawa Shigeyoshi** and Kitamura Tatsuya***

*Department of System Engineering, Graduate School of Science and Engineering,
**Faculty of Information,
Shizuoka University, Japan
[bat,kitazawa,kitamura]@cs.inf.shizuoka.ac.jp

Abstract

Our study focuses on the development of a new feature driven speech-processing strategy for cochlear implant system. In each cycle of stimulation of the cochlea, an electrode, corresponding to second formant frequency was chosen among 14 basilar electrodes. On the base of voiced/unvoiced decision of the speech, an electrode corresponding to first formant frequency was selected for stimulation among 6 apex electrodes. Additionally 4-6 electrodes were chosen for stimulation by maxima energy criteria. Speech intelligibility tests on multi syllable Japanese words within normal hearing listeners by acoustic simulation were provided to evaluate performance of the proposed strategy.
Key words: Cochlear implant system, speech feature, acoustic simulation, speech intelligibility test.

1. Introduction

The purpose of cochlear implant is to generate hearing sensations by electric stimulation the auditory nerve in the cochlea. Several multi channel implant devices have been developed during last 20 years. All signal processing strategies have been used in cochlear implant devices could be divided into three main categories: feature extraction strategies M of N wave form strategies, and hybrid strategies (combination of the feature extraction and the waveform strategies).

The feature extraction strategies deliver information based on speech features such as fundamental and formant frequencies, estimated by implementing various speech processing algorithms. For example, the efficient feature extraction speech processing MultiPEAK strategy [1] extracted the first two formants of the speech signal, along with the voice pitch. In addition to the feature extractor, three band-pass filters were provided to present high frequency information in the speech signal.

In contrast, the waveform strategies transmit some type of waveform derived by filtering the speech signal into different frequency bands on the base of some criteria. In the SPEAK [2] speech processing strategy, up to 10 frequency bands were selected according to spectral maxima criteria. The Continuous Interleaved Sampling (CIS)[3] strategy delivers information extracted from filter bank of N band-pass filters to the implanted electrodes in a non-overlapping fashion, that is, in a way such that only one electrode is stimulated at a time.

The hybrid strategies try to present more information by combination of different techniques including the feature extraction and the waveform strategies.

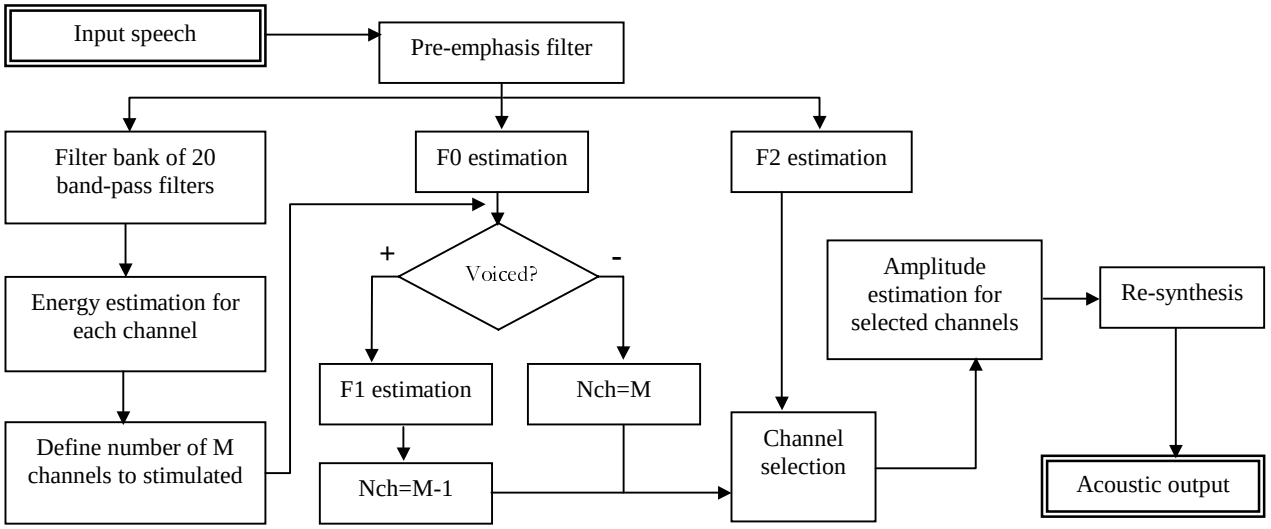

Figure 1. Block diagram of the FDP strategy

2. The proposed strategy

On the base of our previous study [4], we propose a new feature extraction pulsetile strategy for cochlear implant speech processors, named FDP (Feature Driven Pulsetile) strategy. The strategy processes the input speech by combination of feature extraction algorithm with selection of channels within maximum energy. Within the FDP strategy, the input speech was preprocessed through a pre-emphasis filter with cutoff 2000 Hz, and then filtered into 20 frequency bands, using fifth-order Chebyshev type filter with the stop-band ripple 30 dB. Amplitudes for stimulation were calculated from 4ms intervals and then logarithmically compressed. Impulses were delivered up to M electrodes ($M \leq 8$), depending on energy level of each channel. One of the key ideas of the FDP strategy is that in each cycle, an electrode corresponding to F2 frequency must be stimulated according to the estimated F2. In addition, in case of voiced speech, an electrode corresponding to the estimated F1 must be stimulated. Preliminary tests on performance of speech recognition within the strategy show good results. Block diagram of the strategy is illustrated in figure 1. Following subsections describe each stage of the FDP strategy.

2.1. Filter bank

The filter bank used in our acoustic model of FDP strategy consists of 5th order 20 band-pass filters. The frequency spacing of the filter channels is linear in 200-Hz steps from 120-Hz center frequency to 1120 Hz, and then exponential (i.e., linear on a log scale) up to 8890 Hz center frequencies. In this section the spacing is by a factor of 1.250 between 1333 Hz and 3782 Hz, and then by factor of 1.1688 between 4338 Hz and 8890 Hz. Table 1 shows frequency spacing used in our experiments.

Table 1. The center frequencies and bandwidth spacing used in the acoustic model. Filters are numbered from low to high frequency order.

No	Cf.	Bw.	No	Cf.	Bw.
1	120	200	11	2483	360
2	320	200	12	2866	405
3	520	200	13	3297	456
4	720	200	14	3782	513
5	920	200	15	4338	599
6	1120	200	16	4989	701
7	1333	225	17	5749	819
8	1572	253	18	6638	957
9	1841	284	19	7676	1119
10	2143	320	20	8890	1308

2.2. Fundamental frequency estimation

Earlier study [5] of multiple-channel approach showed that relevant speech parameters must be extracted so that phonetically important distinctions are maintained. To make exact decision on voiced or unvoiced region of input speech we extract fundamental frequency.

For estimation of fundamental frequency we have used similar algorithm described in [6]. Output of pre-emphasis filter is low-pass filtered with cutoff frequency 350 Hz, then full wave rectified. The clipping level is determined with 70% of the minimum of the maximum absolute values in the first and last 1/4 samples of the speech segment. Using this clipping level, the speech signal is processed by center clipper. Energy of the frame is calculated from the clipped signal. If it is too small then frame is classified as an unvoiced (silence). Otherwise, the autocorrelation function is computed from the clipped signal. Then the largest peak of the autocorrelation function is located and it is compared to a fixed threshold. In our case threshold was equal to 28% of $R_{(0)}$. If the peak falls below threshold, the segment is classified as unvoiced and if it is above, the location of the peak is defined as the estimated pitch. Figure 2 shows F0 estimation procedure for a voiced segment of a speech signal.

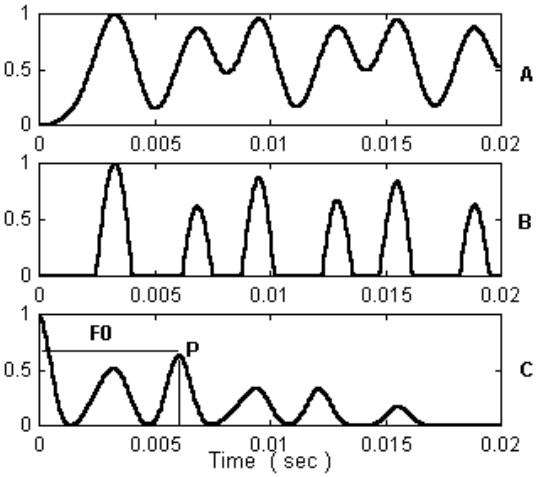

Figure 2. F0 estimation procedure in FDP strategy. (A) Output of low-pass filtered signal, (B) Result of center clipping followed by full-wave rectification, (C) Pitch determination from autocorrelation.

2.3. Formant estimation

Since formants present the immediate source of articulator information, accurate delivery of them to the cochlea is necessary. In our last experiments, we have used the LPC based formant extractor. For the LPC based formant extractors, there are, in general two ways to proceed. One is to compute the roots of A(z),

$$A(z) = \frac{1}{H(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k}, \quad (1)$$

by means of any standard complex root-finding program; the other is to find local maxima in the spectrum envelope derived from the predictor (peak-picking)[7]. We have used the peak picking method to estimate the formants F1 and F2. This method extracts formant information from the predictor coefficients by calculating spectrum envelope of H(z). The formant frequencies F1 and F2 are found by searching the

envelope for local maxima. Figure 3 shows tracks of F1 and F2, estimated by using the formant extractor.

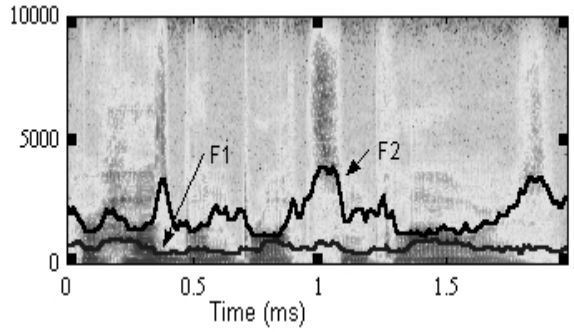

Figure 3. Formant tracks for sentence "Her husband brought some flowers".

2.4. Channel selection

The main difference of the FDP strategy from other strategies is the channel selection algorithm. For a voiced sound two channels are selected according to estimated formant information. Additionally M-1 channels with more energy are selected for stimulation. So for voiced sounds, it is possible to be selected one or more apex channels. In case of unvoiced sound, one channel is selected according to the estimated F2.

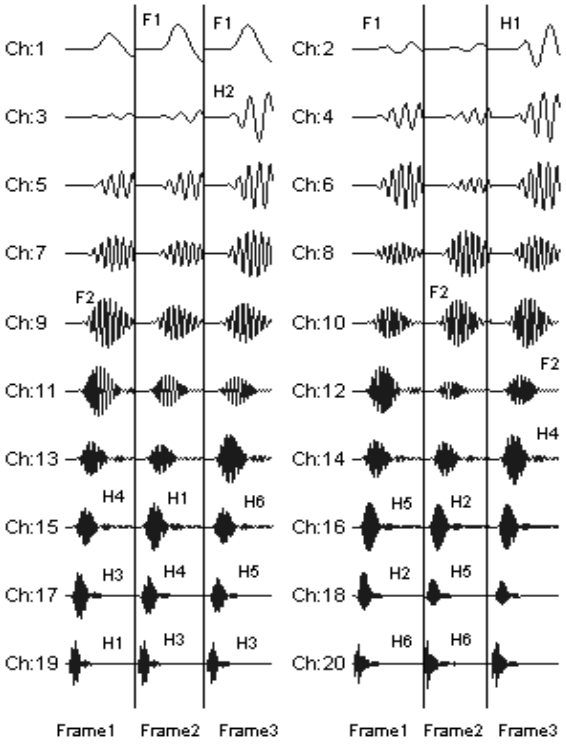

Figure 4. . Example of channel selection within the FDP strategy. Speech from three frames (frame length is 4 ms). Channels are numbered in tonotopic order, while stimulation occurs in order from basilar to apex sequence.

In most case it corresponds to one of basilar electrodes. Additionally M channels are selected by **high-energy** criteria. Figure 4 shows results of channel selection within successive 3 cycles of stimulation process. In the first frame, formant frequency F1 mapped to channel-2, F2 mapped to channel-9, and six more channels were chosen for stimulation: 19^{th} , 18^{th} ,17^{th} , 15^{th} ,16^{th} and 20^{th} . In the second frame, F1 mapped to channel-1, F2 mapped to channel-10, and channels with higher energy were 15^{th} ,16^{th} ,19^{th} , 17^{th} ,18^{th} and 20^{th} . In the last frame F1 mapped to 1^{st} , F2 mapped to 12^{th}, and channels selected by higher-energy criteria were 2^{nd} ,3^{rd} ,19^{th} , 14^{th} ,17^{th} and 15^{th} . Channels 4^{th} ,5^{th} ,6^{th} ,7^{th} ,8^{th} ,11^{th} and 13^{th} were not stimulated within those three cycles.

2.5. Amplitude estimation and re-synthesis

Amplitude estimates for formant information delivery were obtained like in [8], by taking the outputs of the formant filters, passing these through peak detectors, and then finally smoothing the result with a 35-Hz low-pass filter. Envelope amplitudes for filters corresponding to selected channels were computed from output of bandpass filters by full wave rectification and low-pass filtering.

Noise bands were generated from impulse responses of selected filters by filter bank summation method. Impulse trains were estimated from speech of 7ms with frame rate equal to four ms. We don't do any adjustment to improve quality (naturalness) of re-synthesized speech. An example of the acoustic stimuli generated using the strategy is shown in figure 5b.

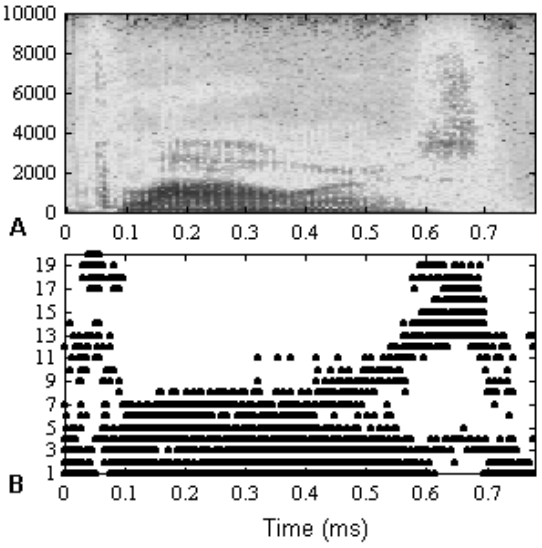

Figure 5. Example with word "flowers". (A) Spectrogram of original utterance, (B) Output of acoustic model of the FDP.

3. Experiment

3.1. Speech material

The aim of this experiment was to evaluate consonant identification within the FDP strategy. The speech test set consisted of 19 /aCv/ syllables with 9 medial consonants

/k,s,n,v,sh,j,r,t,g/ and five vowel environments, /a, i, u, e, o/. The speech material was taken from the MMY speaker's utterance of ATR speech database.

3.2. Procedure

Five normal hearing listeners from our laboratory, ranging in age from 22 to 35 years, participated in the experiment. The listeners are native speakers of Japanese except the third listener. All of them do speech related research and familiar with speech tests.

The test was carried out in a computer room using audiovisual software. Stimuli were listened via closed ear-cushion headphone. The listeners were instructed to understand the presented stimulus and write it on their computers. The listener could listen the presented stimulus at most 5 times. Before the test, the listeners were given a short practice session with examples of speech processed through the acoustic model of the FDP strategy.

4. Results and discussion

Results of the speech recognition test in terms of percent correct are shown in Figure 6. The mean scores for word identification ranged from 50 to 90 percent. The mean scores for medial consonant were ranged from 73 to 90 percent, and the mean scores for final vowel were ranged from 60 to 100 percent.

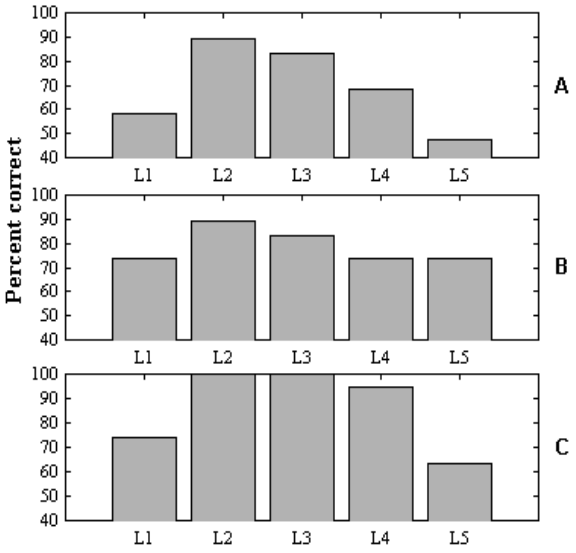

Figure 6. Results of speech recognition test. (A) Word recognition score, (B) Medial consonant identification score, (C) Last vowel recognition score. L1-L5 are notations for listeners.

All listeners show good results in the test. We found that normal hearing listeners could improve their performance score within extended practice of hearing speech signals processed into a small number of channels, because of the non-native Japanese listener (L3) shows stable high score compare to other 3 listeners. Another factor for improving of the performance score on speech tests within normal-hearing listeners is to improve the quality of acoustic stimuli in stage of re-synthesizing. However, here a very question is how much we have to improve if we don't know the exact mechanism of transduction of electrical stimulation in different places of the cochlea in order to achieve natural sounding in the brain.

5. Conclusion

Our study predicts that high level of speech understanding could be attained within the proposed strategy. Selection of electrodes to be stimulated within the proposed algorithm allows appropriate delivery of spectral information to the cochlea. We conclude that summation of noise bands without accurate phase estimation is not sufficient method of representation of electrical stimulation. Another thing is that synthesizing of speech segments from 4ms speech was not enough to re-synthesis voiced segments. So our future studies will concern on accurate estimation of stimulation pulses from adjacent frames of voiced region.

6. References

[1] Patrick, J. and Clark, G., "The Nucleus 22-channel cochlear implant system," Ear and hearing , vol. 12, pp. 3-9, (Suppl. 1), 1991.

[2] Seligman P., McDermott, H.,"Architecture of the spectra 22 speech processor", Annals of Otology, Rhinology and Laringology, (Suppl. 166), 139-141, 1995.

[3] Loizou, P., "Mimicking the ear: An overview of signal processing strategies for cochlear prosthesis", IEEE Signal Processing Magazine, 15(5), 101-130, 1998.

[4] Erdenebat, D., Kitamura, T., and Kitazawa, S."Parametric Study of a Speech Feature Driven Cochlear Implant Speech-Processing Strategy", Proceedings of ASJ,601-602, Spring, 2001.

[5] Blamey, P. et al. "Speech processing studies using an acoustic model of a multiple-channel cochlear implant", J. Acoust. Soc. Am. 76(1), pp. 104-110, 1984.

[6] Dubnowski, J. J., Schafer, R. W., and Rabiner, L. R., "Real-Time digital hardware pitch detector", IEEE Trans. Acoust., Speech, and Signal Proc., Vol. ASSP-24, No 1, pp. 2-8 February 1976.

[7] Pisoni, D.B., "Speech perception: some new directions in research and theory ", Journal of Acoustical Society of America, vol. 78(1), pp. 381-388,1985.

[8] P. Blamey, R. Dowell, and G. Clark, "Acoustic parameters measured by a formant estimating speech processor for a multiple-channel cochlear implant", Journal of the Acoustical Society of America, vol. 82, pp. 38-47, 1987.