EXPLOITING VOICING CUES FOR CONTRAST ENHANCED FREQUENCY SHAPING OF SPEECH FOR IMPAIRED LISTENERS

Naomi Harte , Shahab U. Ansari†, Ian Bruce*

Department of Electrical and Computer Engineering,
McMaster University, Hamilton, Ontario L8S 4K1, Canada

ABSTRACT

This paper investigates the use of voicing information as an additional cue in contrast enhanced frequency shaping (CEFS) of speech to improve perception in the hearing impaired. The presented work builds on an existing system combining multiband compression with contrast enhanced frequency shaping (MICEFS) to restore the auditory nerve response of a hearing impaired listener. CEFS can improve the perception of voiced segments. Hence voicing cues are used to differentiate segments for processing. Alternative processing for unvoiced segments is investigated and shown to improve neural representation of unvoiced segments compared to using MICEFS processing alone.

1. INTRODUCTION

Sensorineural hearing loss manifests itself in the hearing impaired in several ways including reduced dynamic range of hearing and reduced frequency selectivity. Psychophysically, these deficits render loss of speech audibility and speech intelligibility to a hearing-impaired person. Multiband compression is used in conventional hearing aids to compensate for the reduction in dynamic range. Efforts to compensate for impaired frequency selectivity have been largely unfruitful. A better understanding of the effects of hearing loss on the neural representation of speech is clearly needed.

Previous studies of auditory-nerve (AN) fibers in the damaged ear have shown a loss of sensitivity at the best frequency (BF) and a broadening of tuning curves [1]. As a result of the broadened tuning curves, the neural representation of a speech stimulus is degraded [2] [4]. In the unimpaired ear, AN fibers synchronize their responses to the formants of a speech stimulus. This narrowband response of the fibers, referred to as synchrony capture, may be important in perceiving different voiced sounds [3].

Recent years have seen greater interest in the use of physiologically motivated speech processing techniques in an attempt to restore normal (or nearer to normal) AN responses. This paper builds on one such approach – the use of Contrast Enhanced Frequency Shaping (CEFS). Original work in this area by Miller et al [5] was successfully extended to incorporate multiband compression by Bruce [6]. The current work builds on these contributions with enhancements to the algorithm and moves the system closer to a practical implementation in hearing aids. CEFS, as will be detailed in following sections, relies on the knowledge of the formant values in the speech being processed. Much of the previous development work on CEFS has used synthesized vowels with known formants or formant values extracted a priori to CEFS processing. For real-time operation, CEFS processing requires a formant tracker to operate in parallel. This paper considers the practical limitations in availability of accurate formant values and the resultant impact on performance. This work also exploits additional information on voicing available from the formant tracker and looks to take account of voiced/unvoiced cues for CEFS applied to continuous speech.

2. CEFS

CEFS was proposed by Miller et al. [5] in a study exploring the use of spectral modifications to restore the representation of the vowel /ɛ/ in an impaired auditory nerve.

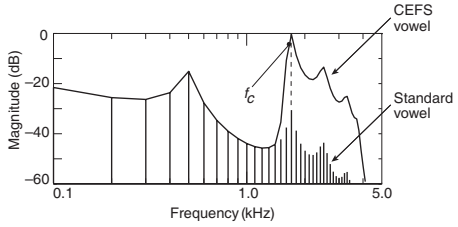


Fig. 1. Line spectrum of vowel /ɛ/ with CEFS modified envelope. f_c is the cutoff frequency of the time-varying highpass filter used in CEFS. Source: Miller et al [5]

The scheme used a time-varying highpass filter with cutoff frequency equal to 50 Hz below F2 of the vowel. The highpass filter applied gain only to F2 and higher frequencies without amplifying harmonics between F1 and F2 as shown in

*Now a Research Fellow at Trinity College Dublin, Ireland. Email: nharte@tcd.ie

†Now a Doctoral candidate at Simon Fraser University, BC, Canada.

This research was supported by NSERC Discovery Grant 261736 and the Barber-Gennum Endowed Chair in Information Technology.

Figure 1. The results for a CEFS modified vowel showed that the AN response to F1 was localized, and the narrowband AN response to F2 was improved. When CEFS modified vowels differing only in F2 were presented to the defective cochlea, the fibers' responses showed changes in rate and in cochlear place with changes in F2 frequency. However, the response to F3 was not restored because of the upward spread of synchrony to F2, and the amplification of harmonics in the trough between F2 and F3. The approach was considered promising overall due to good improvements in the neural representation of stimuli [5].

Bruce subsequently showed that multiband compression could be combined with CEFS without loss of performance [6], unlike other spectral expansion schemes [7]. The speech was first passed though a multi-band compression stage utilizing an FFT-based filterbank and compressor, and then passed through the CEFS amplification stage, using a time-varying FIR filter. This work also extended testing to synthesized sentences.

More recent work by Ansari et al [8] on CEFS has resulted in an improved algorithm MICEFS. The key contributions of this work have been to prevent the upward spread of the synchrony to F2 by emphasizing F2 and F3 without emphasizing harmonics between F2 and F3 and also to restore synchrony to F3 by applying extra gain at F3 relative to the gain at F2 of the speech signal. Results demonstrate that the average discharge rate of fibers in response to a speech signal can be restored to close to normal.

3. FORMANT TRACKER

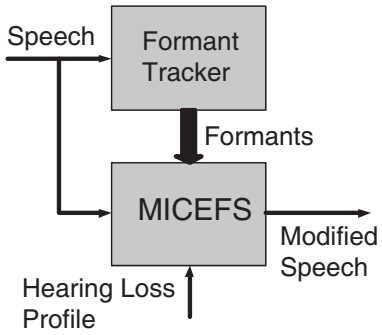


Fig. 2. Real-time MICEFS system.

Figure 2 shows the overall system for MICEFS. The formant tracker used is fully described in [9] and hence only pertinent details are included here. Briefly, the formant tracker is capable of estimating up to the first four formants of voiced speech using a set of time-varying adaptive filters applied over a 20ms LPC window. Within the system, a sample-by-sample decision on whether the preceding 20ms speech segment is voiced or unvoiced is generated. This is derived from a low-frequency to high-frequency energy ratio. The gender of the speaker is automatically tracked and used to modify the cut-off between low and high frequency bands for male and female speakers. Spurious oscillations between voiced and unvoiced are avoided by the use of hysteresis.

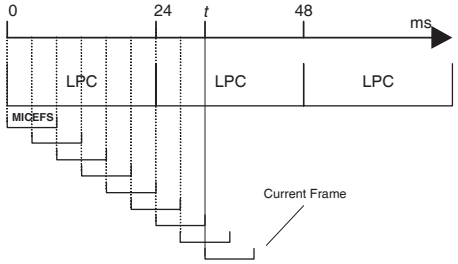


Fig. 3. LPC and MICEFS frame rates in real-time system.

In considering a real-time implementation of MICEFS, it was decided to model the effects of temporal delays from different elements within the system. The formant tracker processes LPC windows of 20ms, whereas the FFT frame size in MICEFS is 8ms with a 4ms overlap. For convenience, the LPC window size was extended to 24ms in this work. The resulting frame processing is shown in Figure 3. For a MICEFs frame at time t as shown, when transforming to the frequency domain, the LPC values for the current frame will not yet be available due to the 10ms delay in calculating formants. This is now modeled in the system. Only the voiced/unvoiced information up to time t will be available in real-time and is used for processing the current frame. The formant tracker operates at 8 kHz, whilst a 16 kHz sampling rate is used in MICEFS. Whilst this is certainly not ideal, only the average formant values for a MICEFS frame are currently utilized and hence interpolation is not an issue.

4. VOICING CUES

The aim of contrast enhanced frequency shaping is to improve perception of voiced segments of speech for hearing aid users. The question hence arises of how to treat unvoiced segments. In previous work on synthesized sentences, MICEFS has been applied to all segments and performance only assessed in terms of formant power ratios (PRs) for the synchronized rates of a population of auditory nerve fibers. At best, the MICEFS processing may not affect the intelligibility of unvoiced sounds, but it is possible that the spectral enhancement, which benefits vowel sounds, may detrimentally affect speech transients such as unvoiced fricatives and plosives. Hence the voiced/unvoiced decision available from the previous frame is allowed influence whether CEFS processing will be used on all speech or only voiced segments. As this information is only available up to the start of the current MICEFS frame, a number of schemes have been investigated to establish the most robust manner in which to utilize the voicing information given this time misalignment. The other

element to the experiments has been to examine the affects of different processing applied to unvoiced segments of speech where voiced segments have MICEFS amplification applied.

5. EXPERIMENTS

5.1. Experimental Set-up

The test sentence used in this paper is the synthesized sentence from a male speaker "Five women played basketball" (courtesy of R McGowan of Sensimetrics Corp, Somerville, MA.). The sentence is phonetically rich, giving a variety of formant trajectories. As a synthesized sentence, the actual formant trajectories were known and used in the analysis of the neural representation of the formants. For MICEFS processing, the formant tracker was used to extract estimates of the formants as discussed.

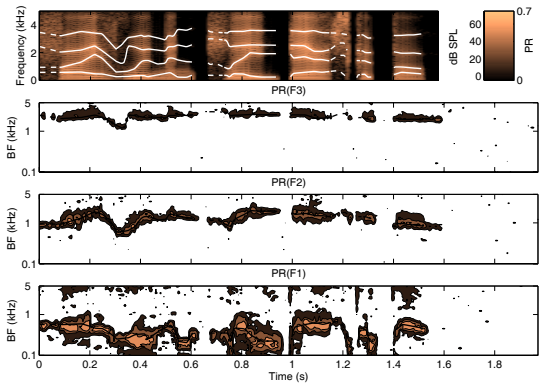


Fig. 4. PR plots for unprocessed sentence presented at 75dB SPL to normal auditory nerve.

The effects of different processing strategies were investigated using the computational model of the auditory periphery as outlined in [10]. For reference, Figure 4 shows the PR (Power Ratio) analysis results for the unprocessed test sentence presented to a normal ear at 75 dB SPL (Sound Pressure Level). These plots show the degree to which a population of model fibers synchronized to the first three formants in the speech spectrum as they change over time.

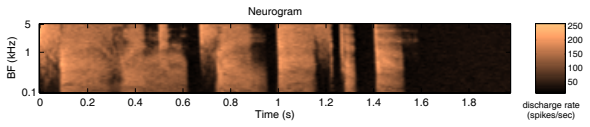


Fig. 5. Neurogram for normal sentence presented at 75dB SPL to impaired auditory nerve.

Figure 5 shows the corresponding neurogram. This shows the short-term average discharge rate of the fibers in response to the sentence as a function of time using a 25.6 ms long Hamming window. The colour bar shows the colour gradients for the average discharge rate of the fibers. Figure 6 shows the PR analysis for the MICEFS processed test sentence presented to an impaired ear at 95 dB SPL. This includes the effects of the real-time formant tracking in the system. Overall, this respesents a superior auditory response from the impaired AN than with unprocessed speech [8]. The neural representation is best for F1 and F2 with some synchrony spread evident in F3.

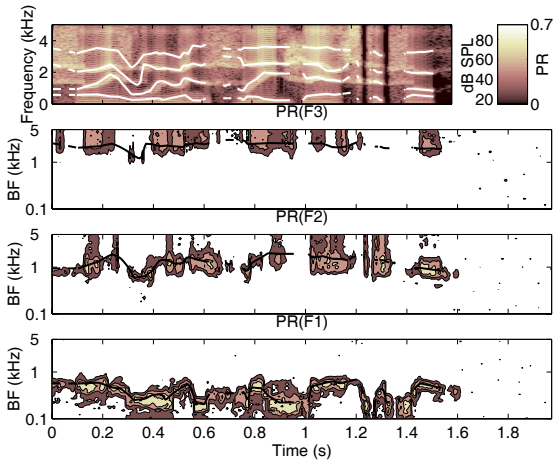


Fig. 6. PR plots for MICEFS processed sentence presented at 95dB SPL to impaired auditory nerve.

The corresponding neurogram for this MICEFS sentence is shown in Figure 7. This demonstrates good restoration of neural activity, especially at high frequencies, when compared to using no MICEFS processing. However, in examing regions of unvoiced speech, it is clear that there is a distortion in the discharge rate versus BF representation of these segments. It is this pattern of voiced to unvoiced transitions that the current work is seeking to restore.

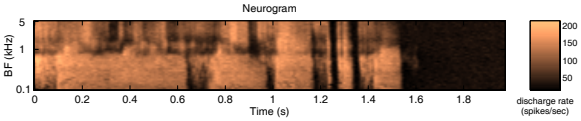


Fig. 7. Neurogram for MICEFS processed sentence presented at 95dB SPL to impaired auditory nerve.

5.2. Voicing Cues

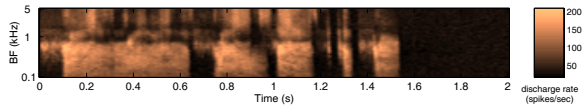
The voicing information was used to change the manner in which unvoiced sections of speech were processed. A number of parameters were varied:

- The percentage of the previous 4ms of speech that had to be identified as unvoiced before the current segment was treated as unvoiced.

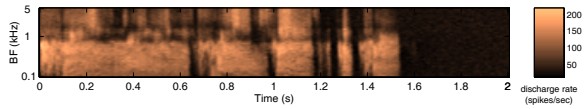
- The filtering applied to voiced segments included NAL-RP, high pass filtering based on hearing profile and prescriptive gain.

Over a large number of experiments it was found that treatment of the current frame as voiced when over 70% of the previous frame was voiced was a sufficiently robust method of using the voicing information from the formant tracker.

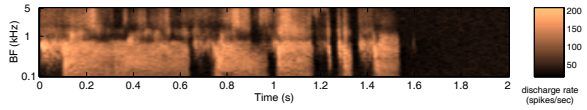
A number of different schemes were invesitgate for the processing of segments of speech classified as unvoiced. This included NAL-RP, NAL-RP with modified prescriptive gain, NAL-RP with added gain in regions of higher spectral energy. The resulting neurograms are shown below. Note that it is only meaningful to examine the neurograms to evaluate any change in the neural representation of the unvoiced segments due to this new processing regime.



(a) NAL-RP



(b) NAL-RP with modified prescriptive gain



(c) NAL-RP with extra high energy gain

Fig. 8. Neurograms with different processing of unvoiced segments

Close inspection of the neurograms in Figure 8 reveals that auditory nerve patterns in unvoiced regions are best reproduced in (c) by using NAL-RP gain in unvoiced regions with additional gain in regions of higher spectral energy. This shows a better restoration of neural representation that in using MICEFS for all speech as shown in Figure 7.

6. CONCLUSIONS

This paper has shown that MICEFs employing real-time tracking of formants can work to restore normal auditory nerve behaviour in voiced segments of speech for the hearing impaired. Different processing has been applied to unvoiced segments in an attempt to maintain lower neural activity in unvoiced segments typically seen in the normal ear response. Some improvements have been found in neural representation in unvoiced segments by using NAL-RP and extra gain in high spectral energy regions. Work is ongoing to establish a quantitative measure of similarity for comparing neurograms and PRs as this work is currently qualitative and prone to error. This will ease extension of testing to a greater database of speech. In the medium term, it is hoped to run trials on human listeners to establish how MICEFS processed speech is perceived in the hearing impaired.

7. REFERENCES

[1] Liberman, M. C. & Mulroy, M. J. “Acute and chronic effects of acoustic trauma: Cochlear pathology and auditory nerve pathophysiology.” In R. P. Hamernik, D. Henderson, & R. Salvi (Eds.), New Perspectives on Noise-Induced Hearing Loss, 105–135 New York, Raven, 1982.

[2] Palmer, A. R. & Moorjani, P. A. “Responses to speech signals in the normal and pathological peripheral auditory system.” Prog. Brain Res., 97, 107–115. 1993

[3] Young, E. D. & Sachs, M. B. “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers.” J. Acoust. Soc. Am., 66, 1381–1403. 1979

[4] Miller, R. L., Schilling, J. R., Franck, K. R., & Young, E. D. “Effects of acoustic trauma on the representation of the vowel /ɛ/ in cat auditory nerve fibres.” J. Acoust. Soc. Am., 101, 3602–3616. 1997

[5] Miller, R. L., Calhoun, B. M., & Young, E. D. “Contrast enhancement improves the representation of /ɛ/ like vowels in the hearing-impaired auditory nerve.” J. Acoust. Soc. Am., 106, 2693–2708. 1999

[6] Bruce, I. C. “Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids”. Physiol. Meas., 25, 945–956. 2004

[7] Franck, B. A., van Kreveld-Bos, C. S., Dreschler, W. A., & Verschuure, H. “Evaluation of spectral enhancement in hearing aids, combined with phonemic compression.” J. Acoust. Soc. Am., 106, 1452–1464. 1999

[8] Ansari, S. U., Bajaj, H., Mustafa, K., & Bruce, I. C. “Time-efficient contrast-enhancing frequency shaping and multiband compression in hearing aids.” International Hearing Aid Conference (IHCON) 2004

[9] Mustafa, K. & Bruce, I. C. “Formant tracking for continuous speech with speaker variability.” IEEE Transactions on Speech and Audio Processing.” (March 2006).

[10] Bruce, I. C., Sachs, M. B., & Young, E. D. “An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses.” J. Acoust. Soc. Am.,113 (1), 369–388. 2003