Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Tonotopic variation in the correlation between the shimmer of a natural vowel and that of the evoked response

Hilmi R Dajani (1), Christian Giguère (1,2)

(1) School of Information Technology and Engineering, University of Ottawa, 161 Louis Pasteur, Ottawa ON, Canada, K1N 6N5
(2) Audiology and Speech-Language Pathology Program, University of Ottawa, 451 Smyth Road, Ottawa ON, Canada K1H 8M5

PACS: 43.64.Bt, 43.64.Qh, 43.64.Ri, 43.64.Sj, 43.71.Qr

ABSTRACT

Measures of amplitude and frequency perturbations in the fundamental frequency (F0) of speech, known as shimmer and jitter respectively, are commonly used to assess speech pathology and voice quality. One limitation of these measures is that they are not based on auditory processing. Shimmer estimation, in particular, could benefit from the incorporation of auditory processing because the outputs of the peripheral auditory filters arranged along the tonotopic axis have very different amplitude modulation profiles at the fundamental periodicity. In this study, we compared the amplitude modulations in the brainstem response evoked by a natural vowel stimulus in seven normal hearing subjects to the shimmer in the broadband stimulus and in the stimulus filtered around each of the first four formants (F1 – F4). The correlation coefficients between the amplitude contour derived from the grand-averaged evoked response and amplitude contours derived from the broadband speech signal and the signal filtered around F1, F2, F3, and F4 were 0.66, 0.35, 0.65, 0.81, and 0.80 respectively. On the other hand, the stimulus amplitude contour variance (a measure of the power of amplitude perturbations) was 20.4, 8.4, 10.1, and 3.8 dB for the unfiltered signal and the signal filtered around F1, F2, and F3 respectively relative to the variance of the amplitude contour of the signal filtered around F4. Therefore, strong correlations with the amplitude contour of the evoked response were obtained for the speech signal filtered around F3 and F4 in spite of having smaller amplitude perturbations compared to the broadband signal and the signal filtered around F1 and F2. This result suggests that shimmer calculated in broadband speech may not be the best measure of perceptually and physiologically relevant amplitude perturbations, and therefore indicates the need for representations that characterize shimmer separately in the different frequency regions of speech.

INTRODUCTION

Cycle-by-cycle amplitude and frequency perturbations in the fundamental frequency (F0) of speech, known as shimmer and jitter respectively, are commonly measured to assess speech pathology and voice quality (e.g. Buder and Strand, 2003). Shimmer and jitter can also reflect the emotional state of the speaker and the social dynamics between the speaker and the listener (Ito, 2004), and may be affected by several pathologies such as such as amyotrophic lateral sclerosis (Aronson et al., 1992), multiple sclerosis (Hartelius et al., 1997), and Parkinson's disease (Li et al., 2008).

One limitation of these two measures is that they do not include a consideration of auditory processing, even though there are several lines of evidence that indicate that the processing and perception of shimmer and jitter varies across the tonotopic axis:

1) Psychophysical and neurophysiological studies have determined that the pitch of a harmonic complex is more salient for lower frequency (resolved) harmonics than for higher frequency (unresolved) harmonics (e.g. Larsen et al., 2008). Other studies have suggested different mechanisms for processing of the pitch of resolved and unresolved harmonics (Carlyon and Shackleton, 1994). While these studies did not directly examine perturbations in the pitch, they suggest that shimmer and jitter may also be processed differently depending on where the acoustic energy is concentrated along the tonotopic axis.

2) Recordings of neural activity in the auditory nerve in animals have shown that the representation of the pitch of complex tones varies as a function of the stimulus frequency content, with the interspike intervals being related to the stimulus waveforms for low frequency stimuli and to the waveform envelope for high frequency stimuli (Cariani et al., 1996). As a consequence, it is likely that processing at the earliest levels of the auditory system would depend on the frequency content of the stimulus signal that is subject to shimmer and jitter.

3) Simulations of how speech is processed by peripheral auditory filters in the region of unresolved harmonics show that although the outputs of these filters are periodic at the fundamental frequency F0, they have very different amplitude profiles depending on the center frequency of the filter (Patterson et al., 1992).

4) A recent study indicates that shimmer in speech that is bandpass filtered around the third formant could be a possible cue for judging the social relationship between Japanese speakers (Ito, 2004).

These results suggest that distinct shimmer cues are available at different points of the tonotopic axis, and consequently that the estimation of shimmer, in particular, could benefit from the incorporation of auditory processing. As a result, in a previous report, we proposed a new reprentation of shimmer, referred to as the "tonotopic shimmer spectral distribution" (Dajani and Giguère, 2009). This representation plots the spectral content of amplitude contours of bandpass filtered speech at different points along the tonotopic axis. An example of this for the vowel /a/ spoken by an adult male is shown in Fig. 1. It shows how the distributions of the spectral content of the amplitude contours differ between the first three formant regions (Fig. 1).

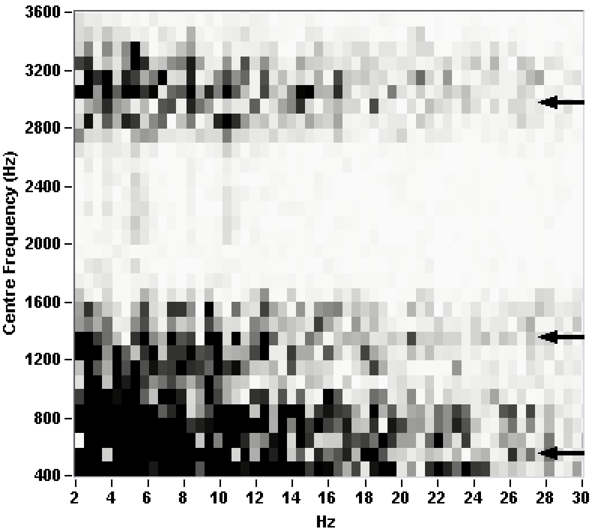

Figure 1. "Tonotopic shimmer spectral distribution" of the vowel /a/ spoken by an adult male (different from the vowel used as the stimulus in the reported experiment). This representation plots the power spectra of shimmer in bandpass filtered speech as a function of the bandpass filter centre frequency (y-axis). The arrows show the first 3 formant frequencies obtained using Praat v.4.5 and averaged over the duration of the vowel.

Brainstem peech-evoked responses present another avenue in which the processing of shimmer can be objectively studied in humans. When measured with a so-called vertical electrode montage, these evoked responses are thought to mainly reflect brainstem activity that is phase-locked to periodicities in the stimulus. We had previously shown that fine structure frequency variations in F0 (i.e. jitter) in a natural vowel can be extracted from speech-evoked responses (Dajani et al., 2005). In that study, the question of amplitude variations at F0 (shimmer) was not addressed. Therefore, in the current report we re-analyse the previously collected evoked responses with the following objectives:

1) To determine if amplitude variations at F0 in the brainstem evoked responses closely track amplitude variations at F0 in the speech acoustic stimulus.

2) To determine the correlation between amplitude variations at F0 in the electrophysiological evoked response and the amplitude variations at F0 in speech stimulus that is bandpass filtered around the first four formants. If different correlation coefficients were found, then this would lend further support to the view that shimmer cues are processed differently and vary in their importance depending on where they originate along the tonotopic axis.

Although the data used for this study had been recorded previously, the objectives, analysis, and results are new and have not been reported before.

METHODS

Subjects and recording of evoked responses

Seven normal hearing subjects (22 – 65 years old, two females) participated in this study. Brainstem speech-evoked potentials were recorded in response to a 2 sec natural /a/ vowel spoken by an adult male using an measurement electrode attached to the scalp at the vertex and a reference electrode placed on the neck just below the hairline. The stimulus vowel had an average F0 of approximately 165 Hz but varied between 162 and 168 Hz over the duration of the utterance. The responses were digitised at 32 kHz and 16 bit resolution. Each experimental session consisted of 1350 stimulus repetitions over which the responses were synchronously averaged to improve the response SNR. A control experiment with the earphone inserted in a Zwislocki coupler confirmed that there was no electrical leakage from the sound generating equipment to the electrodes. Further details regarding the stimulus generation and recording of the evoked responses are found in Dajani et al. (2005).

Analysis

Peak amplitudes in individual pitch periods in the speech stimulus waveform were determined using a semi-automatic method in which a peak detector fits a quadratic polynomial to the signal over successive 6 msec intervals (Fig. 2). The detected peaks were inspected visually, and a few errors were manually corrected using a software program written for this purpose. The peak amplitudes in each pitch period were then interpolated using a cubic spline interpolator at a sampling frequency of 32 kHz. The result of this operation gave a "stimulus pitch amplitude contour" associated with the unfiltered speech signal.

To determine the pitch amplitude contours associated with each of the first four formant regions (F1 to F4), the formant frequencies were first estimated using Praat v.4.5 and averaged over the utterance. Then the speech signal was filtered using a 201 tap bandpass FIR filter with a bandwidth of 400 Hz, and centered on each formant frequency. Given the lack of a good understanding of how the pitch periodicity is represented in the brainstem in different regions of the tonotopic axis, the choice of a bandwidth of 400 Hz is based on a compromise between the need for the filter to be sufficiently narrow to isolate the speech signal around the formant, and to be sufficiently wide to reflect the information that is available in higher centers from a combination of multiple narrow cochlear filters at low tonotopic frequencies. This bandwidth of 400 Hz also corresponds to the bandwidth chosen by Ito (2004) for isolating speech in the region of F3. The output of the bandpass filter was then processed as described above to obtain the associated stimulus pitch amplitude contours, with compensation for the delay of the FIR filter included in the analysis.

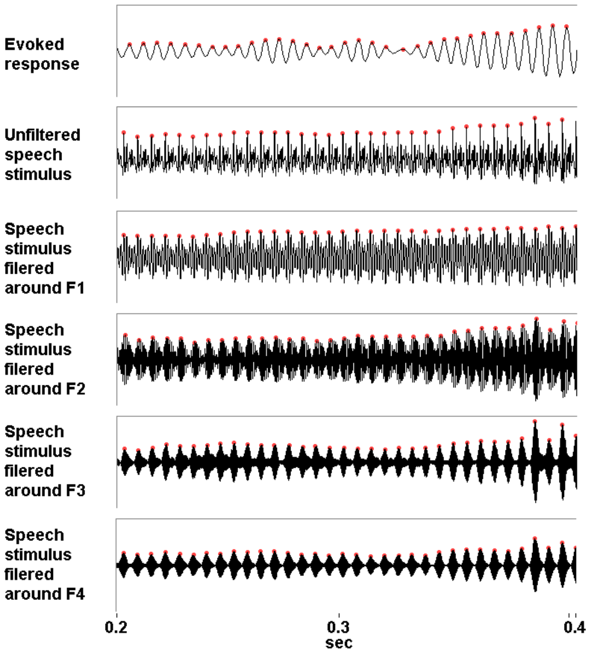

Figure 2. Top to bottom: Detected peaks (red bullets) of individual pitch periods within the first 0.2 second of the evoked response, unfiltered speech stimulus, speech stimulus filtered around F1, filtered around F2, filtered around F3, and filtered around F4. The peaks are used to construct the "pitch amplitude contour" associated with each signal. The signals are shown after compensation for the delay of the FIR filter. For clarity, the scale of the y-axis is adjusted so that the signal fills the plot.

To analyse the evoked response, first the grand-average response was obtained by averaging the responses of all the subjects. Then the signal in the region of F0 was isolated using a 501 tap bandpass FIR filter with a bandwidth of 70 Hz centered at 165 Hz. The output of the filter was processed as described above to obtain the "evoked response pitch amplitude contour", with compensation for the delay of the FIR filter incorporated included in the analysis.

The statistical correlation coefficient between the evoked response pitch amplitude contour and each of the stimulus contours was determined, calculated as the ratio of the covariance between the two curves and the product of the standard deviations of the curves:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Because there is a delay between the stimulus and the evoked response, it was necessary to align the two curves prior to the calculation of ρ. This was done by shifting the stimulus contour forward to the time point where ρ was maximized. In addition to ρ, the variance of each stimulus contour, which is a measure of the power of the amplitude perturbations, was calculated.

RESULTS

Figure 3 shows the pitch amplitude contours obtained from the grand-averaged evoked response, from the unfiltered speech stimulus signal, and from the stimulus signal bandpass filtered in the region of the first four formants. The correlation coefficients between the evoked response contour and each of the stimulus contours is shown in Table 1. The stimulus contour power (or variance) is also shown in units of dB relative to the power of the contour of the signal filtered around F4.

As can be seen, there is a fairly good correlation between the amplitude contour of the unfiltered stimulus and the evoked response (ρ = 0.66). The best correlations, however, were obtained with contours of the stimulus filtered around F3 and F4 (ρ = 0.81 and 0.80). These correlations were observed in spite of having smaller amplitude perturbations relative to the contours of the unfiltered stimulus and the stimuli filtered around F1 and F2. In marked contrast, a much smaller correlation was obtained with the stimulus filtered around F1 (ρ = 0.35) despite having amplitude perturbations 8.4 dB higher than those of the stimulus filtered around F4.

The latency shift required to maximize ρ was between 4 and 5 ms for the unfiltered stimulus, and stimuli filtered around F2, F3, and F4. These latencies (which include a delay of < 1 ms related to the transmission of the stimulus to the earphone) fit with the latencies of 5-10 ms reported for brainstem speech evoked responses (Chandrasekaran and Kraus, 2010). In contrast, the latency of shift required for the stimulus filtered around F1 was 119.9 ms. Since such a latency is too long for a brainstem response, this lends further support to a dissociation between the evoked response pitch amplitude contour and the stimulus filtered around F1.

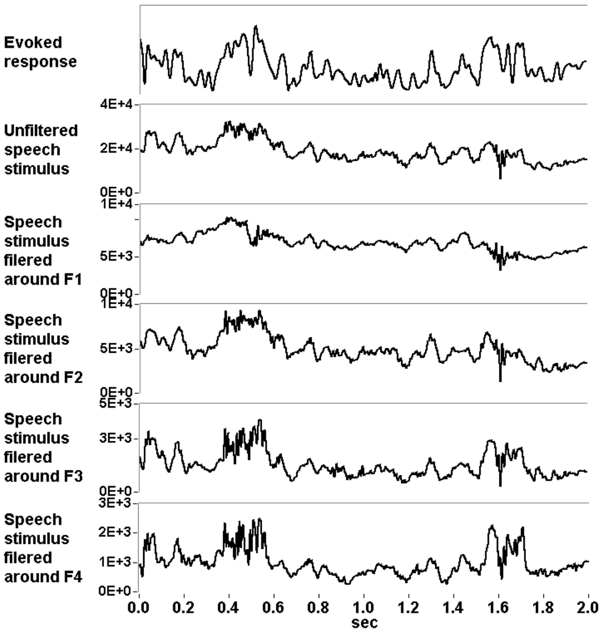

Figure 3. Top to bottom: Pitch amplitude contours for the evoked response, unfiltered speech stimulus, speech stimulus filtered around F1, filtered around F2, filtered around F3, and filtered around F4. The units of the y-axis are arbitrary but the scale is shown for the speech stimuli to illustrate differences in the size of amplitude perturbations.

Table 1. Table shows the correlation coefficient between the pitch amplitude contour of the evoked response and each of the speech stimulus pitch amplitude contours. Also shown is the power (variance) of each of the speech stimulus contours relative to the power of the contour associated with the speech stimulus filtered around F4.

	Correlation coefficient	Relative power (dB)
Unfiltered	0.66	20.4
Filtered around F1 (810 Hz)	0.35	8.4
Filtered around F2 (1491 Hz)	0.65	10.1
Filtered around F3 (2441 Hz)	0.81	3.8
Filtered around F4 (3189 Hz)	0.80	0.0

DISCUSSION AND CONCLUSIONS

This study provides objective physiological evidence in humans that different shimmer cues may be available along the tonotopic axis. The poor correlation between the pitch amplitude contour in the F1 region and the evoked response contour suggests that amplitude perturbations in F0 are processed differently in the various formant regions. This poor correlation may appear surprising, given that the energy around F1 is higher than around the other formants, and given that pitch is known to be more perceptually salient with complexes of resolved (lower frequency) harmonics compared to complexes of unresolved harmonics. However, it is worth distinguishing between the salience of pitch and the perception of perturbation in pitch. Cochlear filters centered at lower frequencies are narrower but have longer time constants. Therefore, it is entirely possible that central mechanisms can extract the pitch associated with resolved harmonics with high saliency, but is unable to track small changes in the pitch well.

How the auditory system processes pitch is still not well understood, but it is possible that pitch coding involves a combination of rate-place activity and the interspike interval distributions (Larsen et al., 2008; Cedolin and Delgutte, 2004). Auditory processing of perturbations in pitch is even less well understood. However regardless of the mechanism involved, this study suggests that shimmer calculated in broadband speech may not be the best measure of perceptually and physiologically relevant amplitude perturbations. It therefore indicates the need for representations that characterize shimmer separately in the different frequency regions of speech.

REFERENCES

A.E. Aronson, L.O. Ramig, W.S. Winholz, S.R. Silber, “Rapid Voice Tremor, or ‘Flutter,’ in Amyotrophic Lateral Sclerosis”, *Annals of Otology, Rhinology & Laryngology*, **101**, 511–518 (1992)

E.H. Buder, E.A. Strand, “Quantitative and Graphic Acoustic Analysis of Phonatory Modulations: The Modulogram”, *Journal of Speech, Language, and Hearing Research*, **46**, 475-490 (2003)

P.A. Cariani, B. Delgutte, “Neural Correlates of the Pitch of Complex Tones I. Pitch and Pitch Salience”, *Journal of Neurophysiology*, **76**, 1698–1716 (1996)

R.P. Carlyon, T.M. Shackleton, “Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?”, *J. Acoust. Soc. Am.* **95**, 3541–3554 (1994)

L. Cedolin, B. Delgutte, “Representations of the Pitch of Complex Tones in Auditory Nerve” in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models* eds. D. Pressnitzer, A. de Cheveigné, S. McAdams, L. Collet (Springer Verlag, New York, 2004) pp. 107-116

B. Chandrasekaran, N. Kraus, “The scalp-recorded brainstem response to speech: Neural origins and plasticity”, *Psychophysiology*, **47**, 236-246 (2010)

H.R. Dajani and C. Giguère, W. Wong, H. Kunov, “Auditory-inspired estimation of jitter and shimmer spectra”, Proceedings of the 16th International Congress on Sound and Vibration, Krakow, Poland, 2009

H.R. Dajani, D. Purcell, W. Wong, H. Kunov, T.W. Picton, “Recording Human Evoked Potentials that Follow the Pitch Contour of a Natural Vowel”, *IEEE Transactions on Biomedical Engineering*, **52**, 1614 –1618 (2005)

L. Hartelius, E.H. Buder, E.A. Strand, “Longterm Phonatory Instability in Individuals with Multiple Sclerosis”, *Journal of Speech, Language, and Hearing Research*, **40**, 1056–1072 (1997)

M. Ito, “Politeness and Voice Quality – The Alternative Method to Measure Aspiration Noise”, Proceedings of Speech Prosody, Nara, Japan, 2004

E. Larsen, L. Cedolin, B. Delgutte, “Pitch Representations in the Auditory Nerve, Two Concurrent Complex Tones”, *Journal of Neurophysiology*, **100**, 1301-1319 (2008)

H. C. Li, H. Dajani, W. Wong, P. Van Lieshout, “Detection of Parksinson’s Voice Using Low Frequency Modulations in Pitch Contour” American Speech-Language Hearing Association Convention, Chicago, U.S.A., 2008 (Poster Presentation)

R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, M.H. Allerhand, “Complex sounds and auditory images”, in *Auditory Physiology and Perception* eds. Y. Cazals, L. Demany, K. Horner (Pergamon Press, Oxford, 1992) pp. 429– 446