Hearing Research 261 (2010) 63–66

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Effects on speech intelligibility of temporal jittering and spectral smearing of the high-frequency components of speech

Ewen N. MacDonald [a,b,c,*], M. Kathleen Pichora-Fuller [d,c], Bruce A. Schneider [d]

[a] Institute of Biomaterials and Biomedical Engineering, University of Toronto, 164 College St., Toronto, Ont., Canada M5S 3G9
[b] Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ont., Canada M5S 3G4
[c] Toronto Rehabilitation Institute, 550 University Ave., Toronto, Ont., Canada M5G 2A2
[d] Department of Psychology, University of Toronto, 3359 Mississauga Road, Mississauga, Ont., Canada L5L 1C6

ARTICLE INFO

Article history:
Received 27 October 2009
Received in revised form 7 January 2010
Accepted 8 January 2010
Available online 20 January 2010

Keywords:
Temporal jitter
Spectral smearing
Speech intelligibility
Word identification

ABSTRACT

In a previous study, we demonstrated that word recognition performance was reduced when the low-frequency components of speech (0–1.2 kHz) were distorted by temporal jittering, but not when they were distorted by spectral smearing (Pichora-Fuller et al., 2007). Temporal jittering distorts the fine structure of the speech signal with negligible alteration of either its long-term spectral or amplitude envelope characteristics. Spectral smearing simulates the effects of broadened auditory filters that occur with cochlear hearing loss (Baer and Moore, 1993). In the present study, the high-frequency components of speech (1.2–7 kHz) were distorted with jittering and smearing. Word recognition in noise for both distortion conditions was poorer than in the intact condition. However, unlike our previous study, no significant difference was found in word recognition performance in the two distorted conditions. Whereas temporal distortion seems to have a deleterious effect that cannot be attributed to spectral distortion when only the lower frequencies are distorted, when the higher frequencies are distorted both temporal and spectral distortion reduce speech intelligibility.

1. Introduction

A speech signal carries information in both the time and frequency domain. Thus a listener may use both temporal and spectral processing to recover the speech information depending on the listening environment. Traditionally, research into the effects of hearing loss on temporal processing has focussed on segmental and supra-segmental processing such as gap-detection, gap-duration discrimination, and temporal-masking paradigms (see Reed et al., 2009 for a review). Recently, much work has been done investigating the role of temporal fine structure, sub-segmental temporal cues, in speech processing (see Moore, 2008 for a review). Processing of temporal fine structure is adversely affected by cochlear hearing loss (Lorenzi et al., 2006; Hopkins et al., 2008) and by age-related losses in synchrony coding (for a review see Pichora-Fuller and MacDonald, 2008) and can disrupt speech intelligibility in noise (Pichora-Fuller et al., 2007). These fine-structure temporal processing deficits have been linked to speech perception in auditory neuropathy (Zeng et al., 1999, 2005b) and linked to temporal and spectral cue trade-offs as well as their relative contributions to speech perception (Drullman, 1995; Smith et al., 2002; Zeng et al., 2004, 2005a).

In a previous study, we explored the effect on speech intelligibility of temporal jittering (Pichora-Fuller et al., 2007). Temporal jittering was chosen in an attempt to simulate the effects on speech intelligibility that might arise from an age-related loss of synchrony coding in the auditory system. The jittering algorithm disrupts the temporal fine structure of a signal by applying a small random temporal jitter to a signal. As an unintended consequence, the jittering algorithm also introduces spectral distortion (i.e., a change in the power spectrum). Insofar as this unintended spectral distortion may have an effect on speech intelligibility in noise, we created a control condition with a similar degree of spectral distortion using a slightly modified version of the smearing algorithm of Baer and Moore (1993, 1994). The results of the speech intelligibility experiments using the smearing algorithm conducted by Baer and Moore (1993, 1994) showed a decrease in word identification when speech was smeared. Their results are similar to previously

Abbreviations: ANOVA, analysis of variance; BW, bandwidth; dB HL, decibels hearing level; dB SNR, signal-to-noise ratio in dB; dB SPL, decibels sound pressure level; D-to-A, digital to analog; FFT, Fast Fourier Transform; Hz, hertz; IFFT, Inverse Fast Fourier Transform; kHz, kilohertz; *M*, mean; ms, millisecond; RAU, rational arcsine units; RMS, root mean square; *SD*, standard deviation; SNR, signal-to-noise ratio; SPIN-R, Revised Speech Perception in Noise Test; STEP, Spectro-Temporal Excitation Pattern

* Corresponding author. Address: Department of Psychology, Queen's University, Humphrey Hall, 62 Arch St., Kingston Ont., Canada K7L 3N6.
E-mail addresses: ewen.macdonald@utoronto.ca (E.N. MacDonald), k.pichora.fuller@utoronto.ca (M.K. Pichora-Fuller).

0378-5955/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.heares.2010.01.005

published results using similar algorithms (Summers, 1991; ter Keurs et al., 1992, 1993; Gnansia et al., 2009). However, in our previous study (Pichora-Fuller et al., 2007), no decrease in word identification was found for the smearing condition. A potential explanation for this discrepancy is that the spectral distortion introduced by the smearing algorithm used in our previous experiments was limited to a low-frequency band (0–1.2 kHz). It is likely that the contributions of temporal and spectral coding may vary across frequency.

The purpose of the present study was to examine the relative contributions of temporal and spectral processing in a frequency range higher than that tested by Pichora-Fuller et al. (2007). In the present study, the smearing and jittering algorithms were applied only to a high-frequency band of speech. As in our previous study (Pichora-Fuller et al., 2007), three speech conditions were tested: an intact (i.e., undistorted) set of sentences, a jittered set of sentences, and a smeared set of sentences. While smaller parameter values for the jittering and smearing algorithms were used than in the previous study, we adopted the same strategy by selecting values such that both manipulations produced the same degree of spectral distortion, but differing amounts of temporal distortion. Thus, any difference in word recognition in noise between the two distorted conditions would be attributable to the reduction in temporal fine structure cues.

2. Materials and methods

2.1. Participants

The participants in this experiment were sixteen young adult paid volunteers (M = 22.6 years, SD = 2.0). All spoke English from at least the age of 5 years and had normal hearing thresholds (pure-tone air-conducted thresholds from 250 to 4000 Hz < 20 dB HL). All participants provided informed consent and their rights as participants were protected. The protocol for the study was approved by the institutional ethics review board.

2.2. Apparatus and stimuli

For the intact condition, the sentences of the Revised Speech Perception in Noise Test (SPIN-R; Bilger et al., 1984) were used. The SPIN-R test consists of eight recorded lists of 50 sentences on one channel and accompanying multi-talker babble on a second channel. The sentences and background babble of the SPIN-R test were digitized at a rate of 20 kHz, and presented monaurally (right ear) over TDH-39P earphones using a Tucker-Davis Technologies System II for D-to-A conversion and to control sentence and background levels. Sentences were presented at 70 dB SPL and the babble level was adjusted to achieve the desired SNR conditions of +8 and 0 dB.

2.2.1. Jittering algorithm

With the exception of parameter values and the frequency band to which it was applied, the jittering algorithm used in the present study was identical to that used previously (Pichora-Fuller et al., 2007). Specifically, if $x(t)$ is the intact waveform, the jittered waveform, $y(t)$, is a time-delayed version of $x(t)$ in which the time delay, δ, varies over time such that $y(t) = x[t - \delta(t)]$. There are two factors contributing to the degree of jitter. The first is the distributions of delays that might occur over a long period of time. The second is the rate at which the delay changes with time. The delay, $\delta(t)$, is generated from the amplitude of a low-pass noise. Thus, the standard deviation of $\delta(t)$ is the root mean square (RMS) amplitude of the low-pass noise. The rate at which the delay changes with time is determined by the bandwidth (BW) of the noise which is limited by the high-frequency cut-off. A lower bandwidth results in a slower rate of change in $\delta(t)$.

In the jittering condition, only the high-frequency components of the SPIN-R speech sentences were jittered. To create the jittered stimuli, a Fast Fourier Transform (FFT) was used to separate the speech signal into its component frequencies. The signal was divided into two bands, a lower band (0–1.2 kHz) and an upper band (1.2–7 kHz). Each band was converted back to the time domain using an Inverse Fast Fourier Transform (IFFT). The upper band was jittered (RMS = 0.07 ms and BW = 500 Hz) and the lower band was left intact. Finally, the two bands were recombined. The average RMS amplitudes of the jittered speech signals were normalized to the same level as the intact signals.

2.2.2. Smearing algorithm

The smearing algorithm used in the present study and in our previous study (Pichora-Fuller et al., 2007) is a slightly modified version of the smearing algorithm developed by Baer and Moore (1993, 1994). The time-domain signal is windowed into overlapping frames and the power spectrum of each frame is calculated. The power vector is then smeared by multiplying it with a smearing matrix. Using the smeared power vector and the unmodified phase vector, an IFFT is calculated to create a frame of the smeared time-domain signal. The frames are then combined using an overlap-add method (Allen, 1977). The smearing matrix is based on models of human auditory filters and requires a single parameter, the smearing factor, to simulate broadened auditory filters with various widths. A smearing factor of 3 simulates a tripling in the bandwidth of the auditory filters. In this study, the smearing was applied only to frequencies greater than 1.2 kHz. Thus, the power components were not smeared for frequencies from 0 to 1187.5 Hz; instead, that portion of the smearing matrix was replaced by an equivalent portion of an identity matrix. Other than restricting the frequency band that was distorted (1.2–7 kHz), the procedure of Baer and Moore (1993) was followed. The smeared sentences were then rescaled to have the same average RMS amplitudes as the intact and jittered sentences.

2.2.3. Matching spectral distortion

As in our previous study (Pichora-Fuller et al., 2007), spectral distortion was evaluated using a pixel-to-pixel comparison of spectrograms and Spectro-Temporal Excitation Patterns (STEPs; e.g., Moore, 2003a,b). The absolute value of the pixel-to-pixel difference in power (using linear units) between the spectrograms of a smeared and intact version of a sentence was calculated and then averaged over time. This resulted in the mean absolute power difference between the smeared and intact versions of each sentence as a function of frequency. Likewise spectral distortion introduced by jittering was similarly quantified. The comparisons were repeated using STEPs instead of spectrograms.

The spectral distortion due to jittering and smearing was quantified as described above using both spectrograms and STEPs for all 400 sentences of the Revised Speech Perception in Noise Test (SPIN-R; Bilger et al., 1984). Unfortunately, the windowing and overlap-add operations used in the smearing algorithm added a small amount of spectral distortion at 125 Hz and higher harmonics. Thus, it was not possible to exactly match the spectral distortion in the jittering and smearing conditions. However, with the exception of the distortion at 125 Hz and higher harmonics, the degree and frequency distribution of spectral distortion was well matched using a smearing factor of 3 for smearing and an RMS of 0.07 ms and BW of 500 Hz for jittering. When limiting the analysis to frequencies greater than 1.2 kHz, the ratio of the spectral distortion for smearing vs. jittering was found to be 1.3 dB using spectrograms, and 2.4 dB using STEPs, indicating that smearing introduced slightly more spectral distortion than jittering.

2.3. Procedure

The listener was asked to report the last word of the sentence immediately following its presentation. In half of the sentences in each list, the last word is predictable from the sentence context (e.g., *The wedding banquet was a feast*) and in the other half it is not predictable (e.g., *We could consider the feast*). Each listener was tested at +8 and 0 dB SNR levels for each of the intact, jittered, and smeared speech conditions. All participants completed the conditions in a fixed order during three sessions, with each session lasting 30–45 min. In session one, they heard intact sentences. In session two, half of the participants heard the jittered sentences and the other half heard the smeared sentences. In session three, the participants heard the other type of distorted sentences. The selection of SPIN-R lists differed for each participant such that lists were counterbalanced over conditions.

3. Results

As in our previous study (Pichora-Fuller et al., 2007), word identification was significantly better in high-context than in low-context sentences and it was better in more favorable SNR conditions (see Fig. 1). The effect of SNR was more pronounced when the context was low. Word recognition in the high-frequency smearing and jittering conditions was significantly reduced compared to performance in the intact condition. For the more favorable SNR condition (+8 dB), word recognition was the same in the high-frequency smearing and jittering conditions. However, for the more adverse SNR (0 dB), word recognition scores were worse in the high-frequency jittering condition compared to performance in the high-frequency smearing condition.

This description was confirmed by a repeated measures analysis of variance (ANOVA) conducted after transforming the data into rational arcsine units (RAU; Studebaker, 1985) with context (high, low), SNR (+8, 0), and distortion condition (intact, jittered, and smeared) as within-subjects factors, and presentation order (jittering first and smearing first) as a between-subjects factor. There were significant main effects of context ($F(1, 14) = 851.91$, $p < 0.001$), SNR ($F(1, 14) = 1088.16$, $p < 0.001$), and distortion condition ($F(2, 28) = 16.331$, $p < 0.001$). There was also a significant two-way interaction between distortion condition × SNR ($F(2, 28) = 5.34$, $p = 0.011$). While a significant between-subjects main effect of presentation order was found ($F(1, 14) = 9.99$, $p = 0.007$), no interaction with presentation order was significant. Multiple pair-wise t-test comparisons using Bonferroni correction confirmed that word recognition was better in the intact condition compared to the smearing and jittering conditions ($p < 0.05$). Although the trend in the difference in performance between the smearing and jittering conditions was not significant, the interaction between distortion condition and SNR was significant. This analysis supports the description that for adverse SNR, word recognition is reduced more in high-frequency jittered than in the high-frequency smeared condition.

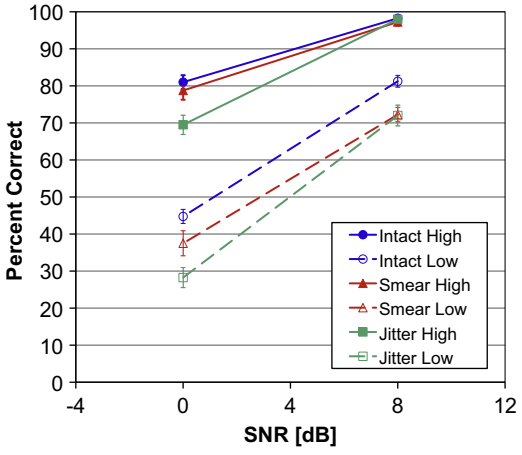

Fig. 1. Word recognition for intact (circles), high-frequency smearing (triangles), and high-frequency jittering (squares) conditions for high- (filled symbols) and low- (open symbols) context SPIN-R sentences. For the high-frequency smearing condition, smearing was applied to the 1.2–7 kHz frequency band using a smearing factor of 3. For the high-frequency jittering condition, jittering was applied to the 1.2–7 kHz frequency band using an RMS of 0.07 ms and BW of 500 Hz.

4. Discussion

The parameters used in the jittering and smearing algorithms in this experiment were different from those used in our previous study (Pichora-Fuller et al., 2007). Specifically, in our previous study, a smearing factor of 6 was used to distort the low-frequency components of speech while in the present study a smearing factor of 3 was used to distort the high-frequency components of speech. Similarly, in our previous study, an RMS of 0.25 ms was used to distort the low-frequency components of speech, while an RMS of 0.07 ms was used in the present experiment to distort the high-frequency components of speech. The reason for choosing a smaller RMS is because the unintended spectral distortion introduced by the jitter algorithm increases with the frequency band of the signal that is being jittered. Thus, for the high-frequency band distorted in the current experiment, the added spectral distortion from the jittering algorithm using an RMS of 0.25 ms (as used in Pichora-Fuller et al., 2007) would be greater than the spectral distortion introduced by the smearing algorithm using a smearing factor of 6. To equate the spectral distortion resulting from jittering and smearing the high-frequency components of speech, either a greater amount of smearing or a lesser amount of jittering had to be chosen. For instance, if the RMS for the jittering were to remain the same then a larger smearing factor would have been required to match the spectral distortion produced by the jittering. Since a smearing factor of 6 was used by Baer and Moore (1993, 1994) to simulate a severe hearing loss, it seemed more appropriate to use a smaller smearing factor because we did not wish to simulate the effects of audiometrically significant hearing loss. We opted to use a smearing factor of 3, used by Baer and Moore to simulate mild to moderate hearing loss, and to reduce the jitter RMS to 0.07 ms so that the amount of spectral distortion would be decreased to a level that would be comparable to that produced by the smearing algorithm. This decision regarding the manner in which to match the distortions, however, makes it difficult to compare the word recognition results of the current experiment to those of our previous study. Nevertheless, the decision allows us to use the results of the present experiment as a bridge between the results of the smearing condition in our previous study, and the results reported from experiments using similar smearing algorithms applied to all speech frequencies (Summers, 1991; ter Keurs et al., 1992, 1993; Baer and Moore, 1993, 1994; Gnansia et al., 2009). One remaining difficulty in directly comparing the recognition scores in the high-frequency smearing condition in the present study to performance in the experiments reported by Baer and Moore (1993, 1994) is that different maskers and speech materials were used across each experiment.

Importantly, unlike the low-frequency smearing condition in our previous study (Pichora-Fuller et al., 2007), a significant decrease in word identification was observed in the present study when smearing was used to distort the high-frequency components of speech. This decrease is also in line with previous published results using similar algorithms (Summers, 1991; ter Keurs et al., 1992, 1993; Baer and Moore, 1993, 1994; Gnansia et al., 2009). When taken together, this combination of results suggests a likelihood that the decrease in word recognition reported in studies using smearing or similar algorithms are mostly due to the

spectral distortion these algorithms introduce in high- rather than low-frequency bands.

For the more favorable SNR, the performance in both high-frequency distortion conditions was similar. For adverse SNR, performance in the high-frequency jittering condition was worse than that of the high-frequency smearing condition. This suggests that there may be differential effects with SNR due to the type of distortion. In easier listening conditions, the main factor contributing to the decrease in performance is the spectral distortion that is common to both high-frequency smearing and jittering. However, as SNR is decreased, the additional temporal distortion in jittering begins to have a significant effect.

For a listener with hearing loss, the processing of both signal and noise is impaired. However, in both the current and the previous study by Pichora-Fuller et al. (2007), the distortions were applied only to the speech signal and not to the babble masker. It is possible that distorting the mixture of speech and noise would lead to further reduction in speech intelligibility beyond that reported here.

In an alternative analysis, psychometric functions were fitted to the low-context results from both the current experiment and the results of our previous study (Pichora-Fuller et al., 2007). The functions fitted were assumed to be cumulative Gaussian distributions with common variance across the conditions of both experiments. Thus, only the threshold, defined as the SNR required for 50% word recognition, was assumed to vary across all the conditions. The thresholds for the intact condition were found to be −1.9 and 0.7 dB in previous and present studies, respectively. Because different listeners participated in each experiment, this small difference is not necessarily surprising. For the low-frequency smearing and jittering conditions, the group thresholds were found to be 0.4 and 4.3 dB below that for the intact condition. For the high-frequency smearing and jittering conditions, the thresholds were found to be 2.3 and 2.8 dB below that of the intact condition. Thus, in comparing the results between the present and previous studies, applying smearing to a high-frequency band of speech is more deleterious to word recognition than applying it to a low-frequency band of speech. This was true even though the smearing factor used in the present study (and applied to a high-frequency band) was half that used in the previous study. This suggests that the decline in speech intelligibility in noise due to broadened auditory filters is associated with difficulty processing higher rather than lower-frequency portions of speech. Nevertheless, the possibility remains that the reduced intelligibility due to distortions affecting primarily the high-frequency components of speech in cochlear pathologies associated with outer hair cell damage may be differentiated from age-related reductions in speech intelligibility in older adults with clinically normal audiograms who may have neural pathologies that primarily affect the temporal processing of the lower-frequency components of speech (e.g., Mills et al., 2006; Caspary et al., 2008).

Acknowledgements

This research was funded by the Natural Sciences and Engineering Research Council of Canada (RGPIN 138572-05; Pichora-Fuller) and the Canadian Institutes of Health Research (MOP-15359; Schneider and Pichora-Fuller). The authors thank Lesley Filmer and Christine DeLuca for assistance in conducting the experiment.

References

Allen, J.B., 1977. Short term spectral analysis, synthesis and modification by discrete Fourier transform. IEEE Trans. Acoust. Signal Process. 25, 235–238.
Baer, T., Moore, B.C.J., 1993. Effects of spectral smearing on the intelligibility of sentences in the presence of noise. J. Acoust. Soc. Am. 94, 1229–1241.
Baer, T., Moore, B.C.J., 1994. Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. J. Acoust. Soc. Am. 95, 2277–2280.
Bilger, R.C., Nuetzel, M.J., Rabinowitz, W.M., Rzeczkowski, C., 1984. Standardization of a test of speech perception in noise. J. Speech Hear. Res. 27, 32–48.
Caspary, D.M., Ling, L., Turner, J.G., Hughes, L.F., 2008. Inhibitory neurotransmission, plasticity and aging in the mammalian central auditory system. J. Exp. Biol. 211, 1781–1791.
Drullman, R., 1995. Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Am. 97 (1), 585–592.
Gnansia, D., Péan, V., Meyer, B., Lorenzi, C., 2009. Effects of spectral smearing and temporal fine structure degradation on speech masking release. J. Acoust. Soc. Am. 125, 4023–4033.
Hopkins, K., Moore, B.C.J., Stone, M.A., 2008. Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. J. Acoust. Soc. Am. 123, 1140–1153.
Lorenzi, C., Gilbert, G., Carn, C., Garnier, S., Moore, B.C.J., 2006. Speech perception problems of the hearing impaired reflect an inability to use temporal fine structure. Proc. Natl. Acad. Sci. USA 103, 18866–18869.
Mills, J.H., Schmiedt, R.A., Schulte, B.A., Dubno, J.R., 2006. Age-related hearing loss: a loss of voltage, not hair cells. Semin. Hear. 27, 228–236.
Moore, B.C.J., 2003a. An Introduction to the Psychology of Hearing, fifth ed. Academic Press, San Diego, CA.
Moore, B.C.J., 2003b. Temporal integration and context effects in hearing. J. Phon. 31, 563–574.
Moore, B.C.J., 2008. The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. J. Assoc. Res. Otolaryngol. 9, 399–406.
Pichora-Fuller, M.K., MacDonald, E., 2008. Auditory temporal processing deficits in older listeners: a review and overview. In: Dau, T., Buchholz, J., Harte, J., Christiansen, T. (Eds.), Auditory Signal Processing in Hearing-Impaired Listeners: Proceedings of the First International Symposium on Audiological and Auditory Research (ISAAR 2007). Centertryk A/S, Denmark, pp. 297–306.
Pichora-Fuller, M.K., Schneider, B.A., MacDonald, E., Pass, H.E., Brown, S., 2007. Temporal jitter disrupts speech intelligibility: a simulation of auditory aging. Hear. Res. 223, 114–121.
Reed, C.M., Braida, L.D., Zurek, P.M., 2009. Review of the literature on temporal resolution in listeners with cochlear hearing impairment: a critical assessment of the role of suprathreshold deficits. Trends Amplif. 13 (1), 4–43.
Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90.
Studebaker, G.A., 1985. A 'rationalized' arcsine transform. J. Speech Hear. Res. 28, 455–462.
Summers, I.R., 1991. Electronically simulated hearing loss and the perception of degraded speech. In: Wise, D.L. (Ed.), Bioinstrumentation and Biosensors. Marcel Dekker Inc., New York, pp. 589–610.
ter Keurs, M., Festen, J.M., Plomp, R., 1992. Effect of spectral envelope smearing on speech reception. I. J. Acoust. Soc. Am. 91, 2872–2880.
ter Keurs, M., Festen, J.M., Plomp, R., 1993. Effect of spectral envelope smearing on speech reception. II. J. Acoust. Soc. Am. 93, 1547–1552.
Zeng, F.G., Kong, Y.Y., Michalewski, H.J., Starr, A., 2005a. Perceptual consequences of disrupted auditory nerve activity. J. Neurophysiol. 93 (6), 3050–3063.
Zeng, F.G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.Y., Chen, H., 2004. On the dichotomy in auditory perception between temporal envelope and fine structure cues. J. Acoust. Soc. Am. 116 (3), 1351–1354.
Zeng, F.G., Nie, K., Stickney, G.S., Kong, Y.Y., Vongphoe, M., Bhargave, A., Wei, C., Cao, K., 2005b. Speech recognition with amplitude and frequency modulations. Proc. Natl. Acad. Sci. USA 102 (7), 2293–2298.
Zeng, F.G., Oba, S., Garde, S., Sininger, Y., Starr, A., 1999. Temporal and speech processing deficits in auditory neuropathy. Neuroreport 10 (16), 3429–3435.