

Auditory perception with slowly-varying amplitude and frequency modulations

Fan-Gang Zeng^{1,2}, Kaibao Nie¹, Ginger Stickney¹ and Ying-Yee Kong²

¹ Department of Otolaryngology – Head and Neck Surgery, University of California, Irvine, CA 92697-1275, USA, fzen^g@uci.edu, knie@uci.edu, stickney@uci.edu

² Department of Cognitive Sciences, University of California, Irvine, ykong@uci.edu

1 Introduction

Amplitude modulation (AM) and frequency modulation (FM) are abundant in natural stimuli, including speech, music, and animal communication sounds. Although amplitude and frequency modulations have been extensively studied physiologically and psychophysically (e.g., Riesz 1928; Grinnell 1963; Suga 1964; Gordon and O'Neill 1998), it is still unclear whether and how the auditory system extracts and uses these cues. For example, there is an ongoing debate on whether amplitude modulation is processed via envelope extraction in the temporal domain (Viemeister 1979) or a second filtering process in the spectral domain (Dau, Kollmeier, and Kohlrausch 1997). It is also unsettled whether frequency modulation is processed independently of amplitude modulation via specialized "FM channels" in the auditory system (Kay and Matthews 1972; Regan and Tansley 1979; Moore and Sek 1996), by a common mechanism (Moore and Sek 1995; Saberi and Hafter 1995).

Regardless of the underlying processing mechanisms of amplitude and frequency modulations, both cues have been shown to contribute to speech recognition in quiet laboratory conditions. Remez et al. (1981; 1990) demonstrated that speech could be reliably recognized with three sinusoids that tracked the formant movement, namely frequency modulation. On the other hand, Shannon et al. (1995) demonstrated that speech could also be reliably recognized with primarily temporal envelope cues, namely amplitude modulation. These results have been traditionally taken as an indication of the redundancy of multiple cues in natural speech sounds.

Motivated by how to deliver the fine structure cue to cochlear implants, recent studies have implicated possible independent contributions of amplitude and frequency modulations to auditory perception (e.g., Smith, Delgutte, and Oxenham 2002). We have developed a signal processing strategy that extracts slowly-varying amplitude and frequency modulations from the traditionally defined temporal envelope and fine structure cues, i.e., Hilbert transform. This novel strategy also

provides a platform to test systematically the independent contribution of amplitude and frequency modulations to auditory perception. Our results suggest that, while amplitude modulation provides essential information for speech recognition in quiet, frequency modulation is needed for speech recognition with competing talkers and music perception.

2 Methods and Materials

2.1 Subjects

A total of 26 normal-hearing listeners participated in the study. Five of them participated in the phoneme recognition experiment, 15 subjects consisting of 3 groups of 5 each participated in the sentence recognition experiment, and additional 6 subjects participated in the melody perception experiment. Local IRB approval and informed consent were obtained.

2.2 Stimuli

Phoneme stimuli included 12 /hvd/ vowels spoken by 3 male, 3 female, and 3 girl talkers (Hillenbrand, Getty, Clark, and Wheeler 1995) and 20 /aCa/ consonants by 2 male and 2 female talkers (Shannon, Jensvold, Padilla, Robert, Wang 1999). The noise was speech-spectrum-shaped and was presented at 0 and -5 dB signal-to-noise ratios. Sentence stimuli were 60 IEEE sentences spoken by a male talker. The noise was a competing sentence spoken by a different male talker. Both sentences had the same onset, but the competing sentence was always longer. Melody stimuli included two sets of 12 familiar songs with one set containing the rhythmic cue and the other containing no rhythmic cue. The rhythmic cue was removed by forcing all notes to be 350 ms in duration with a silent period of 150 ms between notes.

2.3 Processing

Figure 1 displays the basic structure of the novel signal processing strategy that analyzes and synthesizes a stimulus according to its AM and FM components. The original stimulus was first filtered into N narrow-bands (N ranging from 1 to 64 in octave steps). Each narrow-band signal was then subjected to separate AM and FM extraction pathways. The AM was derived by half-wave rectification followed by a low-pass filter. The low-pass filter controlled the amplitude modulation rate, which was set at 500 Hz in the present study. Similar to earlier work on phase vocoders (Flanagan and Golden 1966), the FM was derived by phase-orthogonal demodulators to remove the center frequency of the narrow-band signal. Two independent low-pass filters were used to control the FM depth and rate. In this study, the FM depth was set at 500 Hz or the critical bandwidth, whichever was narrower, while the FM rate was set at 400 Hz. The delay difference was compensated between the AM and FM pathways before recovering the center frequency to re-synthesize the original stimulus.

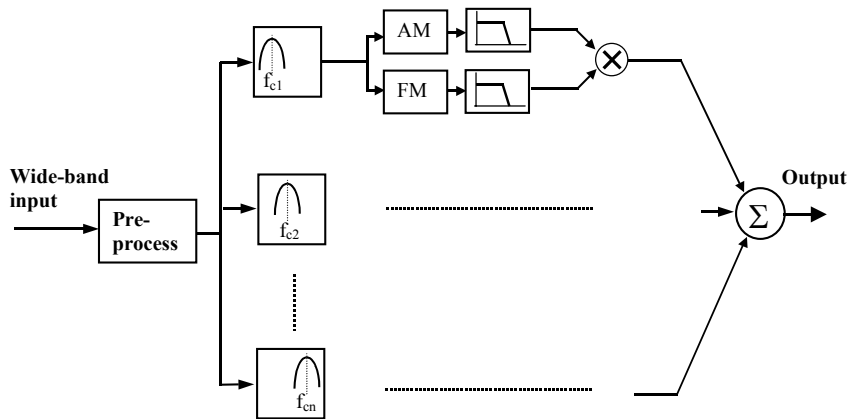


Fig. 1. Signal processing in sound analysis and synthesis using AM and FM cues.

To appreciate the novelty of the proposed processing, a synthetic token /bai/ was used in an 8-band processor to contain AM only and AM+FM components. Figure 2 shows the spectrogram of the original token (left panel), the AM only token (middle panel), and the AM+FM token (right panel). Although neither the AM only nor AM+FM token contained the detailed harmonic structure as in the original token, the AM+FM token clearly preserved formant transition information (see the initial formant transitions from /b/ to /a/ in the first 40 msec of the stimulus as well as the much longer transitions from /a/ to /i/ for the last 300 msec of the stimulus).

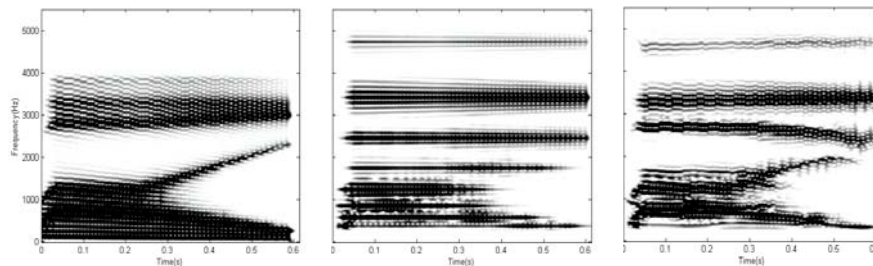


Fig. 2. Spectrograms of the original token /bai/ (left panel), the 8-band AM only token (middle panel), and the 8-band AM+FM token (right panel).

2.4 Procedures

In the phoneme recognition experiment, the subject was asked to identify the randomly presented phoneme by clicking on the GUI that contained all possible phonemes. Trial-by-trial feedback was provided. In the sentence recognition experiment, the subject heard 60 target sentences both in quiet and in the presence of a single competing sentence at different signal-to-noise ratios. The subject was then asked to type in the sentence via a keyboard. No feedback of any form was given. The keywords correctly identified were computed and reported as percent

correct. In the melody recognition experiment, the subject heard a melody and had to choose from 1 of the 12 melodies whose names were displayed on a computer screen. Trial-by-trial feedback was provided. A practice run was always given before formal data collection. All stimuli were presented monaurally through a Sennheiser headphone at 65 dB SPL. The subject performed these experiments in a double-walled, sound-attenuated chamber.

3 Results

3.1 Phoneme recognition

Figure 3 shows vowel (left panel) and consonant (right panel) recognition scores as a function of signal to noise ratios in the 8-band condition. The vowel recognition was generally at a high level between 70 and 90% correct for all conditions. A repeated measures ANOVA revealed no statistical difference between the AM and the AM+FM conditions [$F(1,4)= 4.748, p=0.095$] but a significant difference between the noise and quiet conditions [$F(2,8)=38.663, p<0.001$]. On the other hand, a significant difference in consonant recognition was observed for both the modulation [$F(1,4)= 35.911, p=0.004$] and noise [$F(2,8)= 114.534, p<0.001$] factors. A significant interaction was also observed between the modulation and noise factors, with the AM+FM stimuli producing better performance than the AM only stimuli in noise but essentially no difference in quiet.

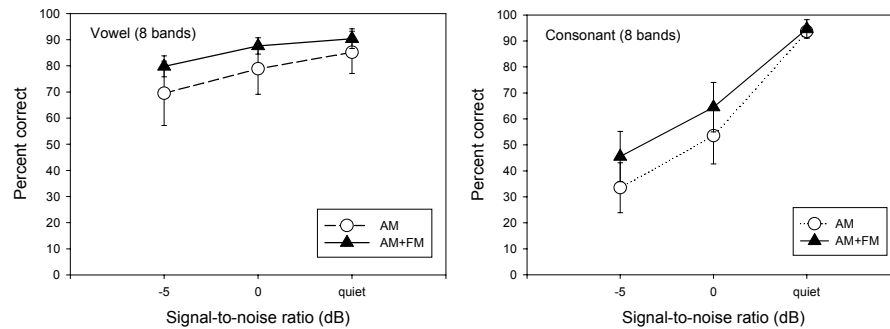


Fig. 3. Vowel (left panel) and consonant (right panel) recognition as a function of signal-to-noise ratios in an 8-band processor. The open circles represent data collected with the AM only condition while the filled triangles represent the AM+FM condition.

3.2 Sentence recognition

Figure 4 shows sentence recognition scores as a function of signal-to-noise ratios in the presence of a competing talker. Different from the modest difference in phoneme recognition, the additional FM cue produced significantly better results

than the AM only condition [$F(2,24) = 452.08, p < .001$], particularly at low signal-to-noise ratios where the improvement was as much as 70 percentage points.

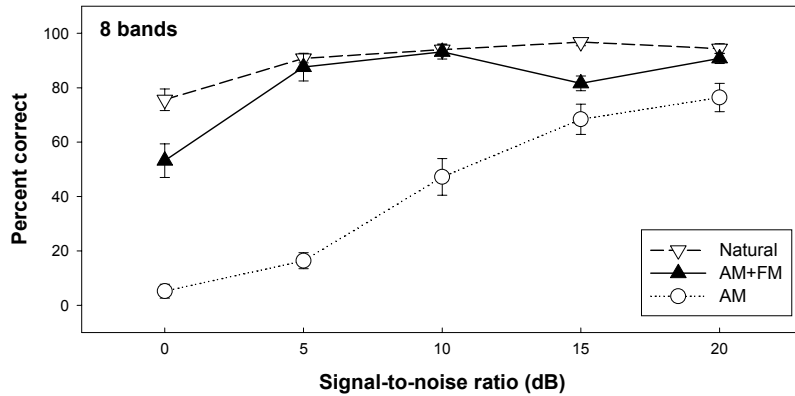


Fig. 4. Sentence recognition as a function of signal-to-noise ratios in an 8-band processor. The noise was a sentence from another talker.

3.3 Music perception

Figure 5 shows melody recognition as a function of the number of frequency bands with (left panel) and without (right panel) the rhythmic cue. Clearly the rhythmic cue contributed to a relatively high level of performance independent of both the number of bands and the addition of the FM cue. However, when the rhythmic cue

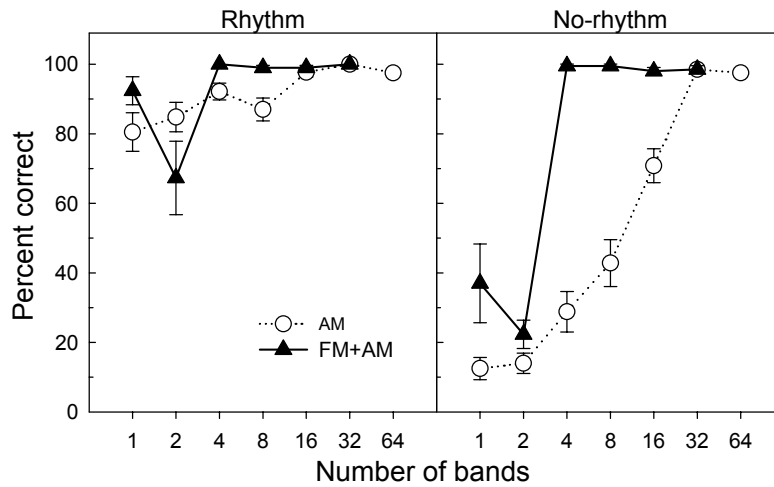


Fig. 5. Melody recognition as a function of the number of bands in the presence (left panel) and absence (right panel) of the rhythmic cue.

was removed, the AM condition needed 32 bands of spectral information to achieve perfect melody recognition while the AM+FM condition only required 4 bands. Between 4 and 16 bands, the AM+FM condition produced significantly better performance (t tests, $P < 0.01$) than the AM condition. The dip in performance with the 2-band FM condition was possibly due to an inappropriate FM representation of the original melody information.

4 Discussion

Together with previous studies, the present data show that, while AM information is sufficient for speech recognition in quiet, FM information is required for speech recognition in noise and for melody recognition without rhythmic cues. The largest improvement was observed for sentence recognition with a competing talker, emphasizing the importance of the FM cue in speech perception under realistic listening environments, e.g., at a cocktail party. We have collected preliminary data suggesting that the FM cue might have allowed the listener to tell one talker (signal) apart from the other (noise). In other words, there appears to be an independent contribution of the AM and FM cues to speech recognition: the AM mostly contributes to “what is said” whereas the FM mostly contributes to “who says what”.

Different from previous studies in which FM might consist of rapid changes across multiple critical bands (Remez, Rubin, Pisoni, and Carrell 1981), the present study only extracts the slowly-varying FM components around the center frequency of a frequency band. With both the modulation depth and rate limited to a few hundred Hertz, this slowly-varying FM cue might be used by cochlear implant users as an efficient means for encoding fine structure information.

The basic principle underlying the AM and FM cues may also be applied to low-rate, high-quality audio coding and processing. For example, for the 5000-Hz sub-band, there will be no need to transmit the 5000-Hz information, rather an FM signal with a bandwidth of 500-Hz or less is needed for transmission.

5 Summary

We have developed a signal processing strategy that can independently extract slowly-varying amplitude and frequency modulations within a frequency band with the number of bands as an independent variable. While the AM provides sufficient information for speech recognition in quiet, the additional FM significantly improves speech recognition in noise and music perception. The FM may be used as an efficient means to convey the fine structure information in cochlear implants and audio coding.

Acknowledgments

We thank Ackland Jones, Michael Vongphoe, Elsa Del Rio, and Sheetal Desai for technical support. This research was supported by grants from NIH (R01-DC02267

and F32-DC-5900) and Chinese NSF (30000041). To be presented at the 13th International Symposium on Hearing.

References

- Dau, T., Kollmeier, B. and Kohlrausch A. (1997) Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.* 102, 2892-2905.
- Flanagan, J.L. and Golden, R.M. (1966) Phase Vocoder. *Bell Sys. Tech. J.* 45, 1493-1509.
- Gordon, M and O'Neill, W.E. (1998) Temporal processing across frequency channels by FM selective auditory neurons can account for FM rate selectivity. *Hear. Res.* 122, 97-108.
- Grinnell, A.D. (1963) The neurophysiology of audition in bats: Intensity and frequency parameters. *J. Physiol.* 167, 38-66.
- Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K. (1995) Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099-3111.
- Kay, R.H. and Matthews, D.R. (1972) On the existence in the human auditory pathway of channels selectively tuned to the modulation present in frequency-modulated tones. *J. Physiol.* 225, 657-677.
- Moore, B.C.J. and Sek, A. (1995) Effects of carrier frequency, modulation rate, and modulation waveform on the detection of modulation and the discrimination of modulation type (amplitude modulation versus frequency modulation). *J. Acoust. Soc. Am.* 97, 2468-2478.
- Moore, B.C.J. and Sek, A. (1996) Detection of frequency modulation at low modulation rates: evidence for a mechanism based on phase locking. *J. Acoust. Soc. Am.* 100, 2320-2331.
- Regan, D and Tansley, B.W. (1979) Selective adaptation to frequency-modulated tones: Evidence for an information-processing channel selectively sensitive to frequency ranges. *J. Acoust. Soc. Am.* 65, 1249-1257.
- Remez, R.E., Rubin, P.E., Pisoni, D.B. and Carrell, T.D. (1981) Speech perception without traditional speech cues. *Science* 212, 947-949.
- Remez, R. and Rubin, P.E. (1990) On the perception of speech from time-varying acoustic information: contributions of amplitude variation. *Percept. Psychophys.* 48, 313-325.
- Riesz, R.R. (1928) Differential intensity sensitivity of the ear for pure tones. *Phys. Rev.* 31, 867-875.
- Saberi, K. and Hafter, E.R. (1995) A common neural code for frequency- and amplitude-modulated sounds. *Nature* 374, 537-539.
- Shannon, R.V., Jensvold, A., Padilla, M., Robert, M.E. and Wang, X. (1999) Consonant recordings for speech testing. *J. Acoust. Soc. Am.* 106, L71-L74.
- Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. and Ekelid, M. (1995) Speech recognition with primarily temporal cues *Science* 270, 303-304.
- Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87-90.
- Suga, N. (1964) Recovery cycles and responses to frequency modulated tone pulses in auditory neurons of echo-locating bats. *J. Physiol.* 175, 50-80.
- Viemeister, N.F. (1979) Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.* 66, 1364-1380.