Auditory-Visual Speech Recognition with Amplitude and Frequency Modulations

Christopher Groot[1] and Chris Davis[2]
[1] Department of Psychology, The University of Melbourne. [2] MARCS Auditory Laboratories, UWS, Sydney, Australia

cgroot@student.unimelb.edu.au; chris.davis@uws.edu.au

Abstract

A recent study by Zeng et al (2005) [PNAS, 102, 2293-2298] demonstrated the importance of FM cues for auditory speech identification in a competing noise environment. The current speech identification study investigated this finding for both an Auditory Only (AO) and an Auditory-Visual (AV) speech in noise identification task. The results demonstrated an FM advantage (compared to AM only) for both the AO and AV speech conditions. This finding shows the need cochlear implant speech processors to encode both AM and FM information even if visual speech cues are present.

1. Introduction

Speech contains multiple cues that can support its reliable identification. For example, Shannon, Zeng, Kamath, Wygonski and Ekelid (1995) have demonstrated that reliable speech recognition can occur based primarily upon temporal envelop cues (amplitude modulation, henceforth AM). Similarly, using stimuli consisting of three sinusoids tracking formant trajectories, Remez and colleagues (Remez, Rubin, Pisoni & Carrell, 1981; Remez & Ruben, 1990) have shown that speech can also be recognized based primarily on frequency cues. Recently, it has been suggested that AM and FM information may provide independent cues for auditory perception (Smith, Delgutte & Oxenham, 2002). Determining the relative importance of such cues for speech identification is important in the design of the speech processing algorithm of cochlear implants since not all speech information can be encoded.

As mentioned above, experiments by Shannon and colleagues have demonstrated reliable speech identification based upon temporal envelope cues; what is remarkable is that high speech recognition rates can be achieved from filtered speech having as few as four narrow frequency bands (Shannon et al, 1995). Such demonstrations have likely influenced the selection of implant speech processing algorithms as the majority of processing strategies are concerned with the extraction of temporal rather than spectral cues (e.g., Compressed Analogue Simultaneous stimulation; CIS; SPEAK; SAS; PPS; MPD & ACE versus F0F1; MPEAK).

However, although AM cues provide sufficient information for speech recognition in quiet, such cues are extremely susceptible to background environmental noise (e.g., competition speech, see Zeng et al., 2005). Recently, Nie, Zeng and colleagues have developed a novel signal processing technique that uses the Hilbert transform to extract the slow varying amplitude and frequency components from speech frequency bands (Nie, Stickney & Zeng, 2005; Zeng et al, 2005, see Figure 1).

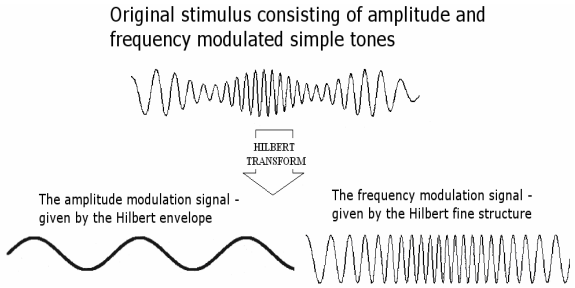

Figure 1. An example of the Hilbert transform that separates a signal into a slowly varying and a rapidly varying component (Modified from Zeng, 2004)

This technique allows for slow varying AM and FM contributions to speech recognition to be independently investigated (the ability to index these slow moving contributions is important since these are most relevant to the implementational constraints of implants). Nie, Zeng and colleagues called their processing algorithm the Frequency-Amplitude-Modulation-Encoding (FAME) strategy. This consists of the following stages. A band-pass filter separates the wide-band signal into n sub-bands. For each sub-band the AM and FM components are extracted. The AM is derived by a rectifier and low-pass filter and the AM and FM pathways are re-synchronized. The FM component is decomposed by two phase-orthogonal demodulators (a

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

cosine and a sine carrier at the centre frequency of the sub-band) and a subsequent low-pass filter. After demodulation, the frequencies of the FM spectrum are shifted from high to low and a narrow-band envelope and its frequency modulation is calculated. This narrow-band envelope is used to threshold FM noise and the signal smoothed by a low-pass filter. The sound signal is re-synthesized by recovering the carrier frequency with the phase signal calculated by using integration. Finally, to form the synthetic sound, the FM signals are modulated by AM signals and summed. An illustrative example of what FAME processing does to speech is shown in Figure 2.

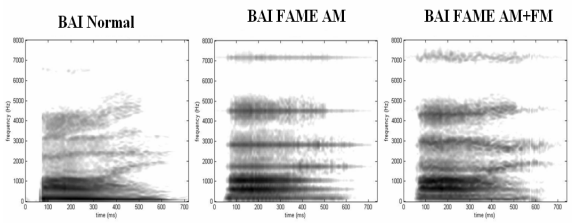

Figure 2. Spectrograms of the unprocessed syllable /bai/ (left panel), 8 band FAME AM processed (middle panel) and 8 band FAME AM+FM processed (right panel). AM+FM has preserved formant transition information.

As an analogue to the limited band information available to cochlear implant uses, Zeng et al (2005) presented to normal-hearing listeners sentences processed with the FAME algorithm which included amplitude modulation only and both amplitude and frequency modulations. These sentences were presented in a quiet and a noise condition (that consisted of competing speech). The results demonstrated the importance of the addition of FM cues for speech recognition in noise (a situation that is more typically encountered in realistic hearing environments). For example, with 8 band filtering, adding the FM component produced an improvement in speech recognition of 23% for sentences presented in silence and a 33% improvement for those in noise (5 dB SNR). Zeng et al suggested that FM provides a prominent cue allowing a listener to separate, and then appropriately assign AM cues in order to form foreground and background auditory objects (important in following a talker against competing speech). This is an important finding as it suggests that FM in addition to AM should be extracted and encoded in cochlear implant speech processing.

However, there is another source of speech related information that has not been taken into account in the Zeng et al (2005) study. It has long been known that the provision of visual speech information (speech associated movements of the lips, mouth, jaw, face and head) enhances the intelligibility of degraded acoustic speech signals (Sumby & Pollack, 1954). It is obvious that visual speech also provides cues to form foreground and background auditory objects. This raises the possibility that FM cues will be redundant if the listener also has access to a talker's visual speech (a situation that will be common and exploited by people who have hearing difficulties). The current study is a modified replication of the Zeng at al study which tests the AM and AM+FM FAME processed speech in quiet and in noise conditions with and without accompanying visual speech.

2. Methods

2.1. Stimuli

Eighty IEEE sentences were selected from the 720 available in the 1969-revised list of phonetically balanced sentences (IEEE/Harvard, 1969). These were selected on the basis of their appropriateness for the age group of participants, as assessed by multiple raters. IEEE sentences were employed as they exhibit complex structure and information, and provide stimuli for which the advantage of semantic and contextual information to aid interpretation is limited (Zeng et al, 2005).

The selected sentences were spoken by a male native English speaker in a double walled, sound attenuated room. Both audio and video were recorded using a Sony DCR-DVD602E Digital DVD Handycam. A commercially available babble track (containing simultaneous multiple male and female speakers – www.auditec.com) was used as competing noise stimuli.

2.1.1. *Speech signal processing*

Audio signal processing was performed using the FAME algorithm used in the Zeng et al study. This encoding strategy creates the band filtered signal via a Hilbert transform, after which the signal is degraded into AM and FM components. Unlike the multiple bands tested in the Zeng at al study, the current study only used 8 frequency bands. To derive the AM information cues, the temporal envelope of each sub-band was modulated by each band's center frequency. The 8 modulated band frequencies were then combined to create the AM audio stimuli. FM information cues were derived by removing the centre frequency in each sub-band after the Hilbert transform and limiting the fine structure of each to the lesser of either: a frequency range of 400Hz to 500Hz, or the Hilbert transform filter's bandwidth. Importantly, before summation of the 8 AM and 8 FM modulated sub-bands to create AM+FM stimuli, the modulated signals were submitted to an additional Hilbert transform band pass filter to eliminate any extra spectral information created as part of the frequency modulation process.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

The audio portion of both the target IEEE sentences and competing 'babble' noise were subjected to the FAME process. The AM and FM modulated competing noise was summated with the AM and FM modulated speech signals at -5dB signal to noise ratio (SNR). For all stimuli containing competing noise, the onset of noise occurred prior to that of speech stimuli, and had a longer duration. For visual speech conditions, after the speech signal was processed, it was dubbed to the video of the talker.

2.2 Participants

Fifty-seven first year undergraduate university students participated in the experiment. Participants were native speakers of English, 18 years of age or over and had no history of reported hearing loss and normal or corrected-to-normal vision.

2.3 Procedure

All participants were tested individually in a separate booth. Auditory stimuli were presented through Sennheiser HD580 headphones. The video clips (compressed to MPEG 2) were played back using the DMDX software (Forster & Forster, 2003) on a ViewSonic G810 21 inch monitor at a resolution of 1024 x 768. In conditions for which there was no visual speech, the participants viewed a blank screen. All participants heard a total of 80 IEEE sentences presented in 8 conditions (i.e., 10 sentences in each condition): 1. Auditory-Only (henceforth AO) AM (in quiet), 2. AO AM+FM (in quiet), 3. AO AM (in noise), 4. AO AM + FM (in noise), 5. Auditory-Visual (AV) AM (in quiet), 6. AV AM + FM (in quiet), 7. AV AM (in noise) and 8. AV AM + FM (in noise). Presentation of sentences within each condition block was randomized, as well as the presentation of the condition blocks themselves. Eight versions of each experiment were prepared such that no item would be repeated in any version but each version would contain all conditions (i.e., sentence for all conditions were fully rotated). The percentage correct word identification scores were calculated as the measure of speech recognition for each condition. In scoring these data, credit was given only if the typed word exactly matched the spoken word (except where the response was an obvious typo). At the commencement of each testing session, participants were provided with a set of practice items to familiarize them with the various conditions of degraded speech. In the experimental phase, participants provided responses to stimuli by keyboard.

3. Results

3.1 Word recognition in quiet.

Mean percentage correct word recognition scores for each condition presented in quiet is presented in Figure 3.

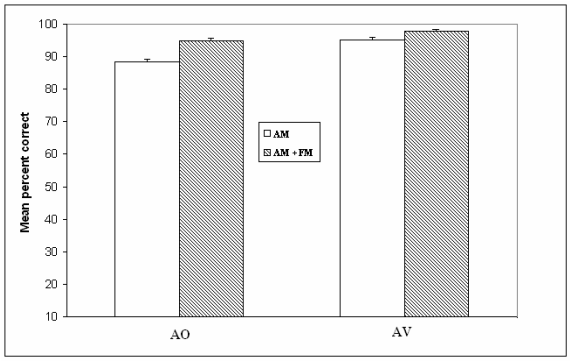

Figure 3. Mean percent correct word recognition scores for each condition presented in quiet. (AM = AM cue; AM+FM = AM + FM cue; AO = Auditory Only; AV = Auditory-Visual Speech).

As can be seen, performance was near ceiling for all conditions. This was particularly the case for the AM+FM conditions. The participant data was analyzed with a 2 (AO versus AV) X 2 (AM versus AM+FM) ANOVA. More words were correctly identified in the AV condition than in the AO condition [$F1\ (1,50) = 43.73$, $p < 0.05$] (a visual enhancement effect). Likewise there were more words correct in the AM+FM condition compared to the AM condition [$F1\ (1,50) = 43.92$, $p < 0.05$] (an FM effect). There was a significant interaction between these effects, such that the FM effect was greater in the AO condition compared to the AV condition, [$F1\ (1,50) = 10.08$, $p < 0.05$]. This was most likely due to a ceiling effect in the AV condition.

3.2 Word recognition in competing noise.

Mean percentage correct word recognition scores for each condition presented in noise is presented in Figure 4.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

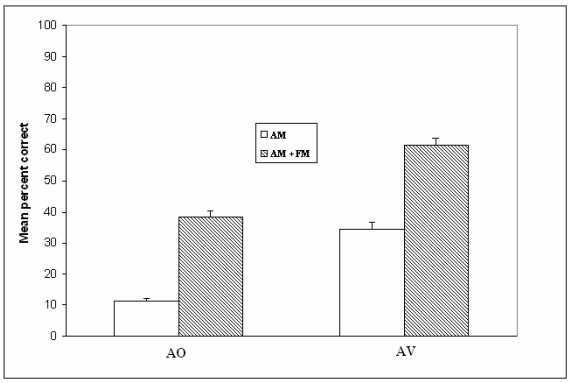

Figure 4. Mean percent correct word recognition scores for each condition presented in noise. (AM = AM cue; AM+FM = AM + FM cue; AO = Auditory Only; AV = Auditory-Visual Speech).

As with the in quiet data, the in noise data was analyzed with a 2 (AO versus AV) X 2 (AM versus AM+FM) ANOVA. More words were correctly identified in the AV condition than in the AO condition [F1 (1,50) = 243.60, $p < 0.05$] (a visual enhancement effect). Likewise there were more words correct in the AM+FM condition compared to the AM condition [F1 (1,50) = 273.13, $p < 0.05$] (an FM effect). However, unlike the in quiet data, there was no interaction between these effects, $F < 1$.

It is clear from comparing Figures 3 and 4 that presenting the sentences mixed with multi-talker babble had a deleterious effect on speech recognition. For instance, similar to Zeng at al there was a considerable drop in correct identification scores for AO AM speech in quiet versus in noise (in this case from 89% correct to 11%). As expected, a comparison of percentage correct in quiet and in noise was significant [F1 (1,50) = 1977.05, $p < 0.05$]. The variable of noise (in quiet versus in noise) also interacted with the other variables, such that there was a greater FM effect in the noise condition, [F1 (1,50) = 134.57, $p < 0.05$] and a greater visual enhancement effect in noise, [F1 (1,50) = 173.97, $p < 0.05$]. However, there was no interaction between the variables of noise; FM effect and visual enhancement interaction, [F1 (1,50) = 1.42, $p > 0.05$].

If the information provided by FM cues was the same as that provided by visual speech cues, then (under a simple additive model of the relationship between speech-related information and identification) there should not have been an FM effect in the AV condition. However, it is clear from the results that this was not the case since in noise condition (where there was no ceiling effect), there was a significant FM effect in the AV condition [F1 (1,50) =127.88, $p < 0.05$]. This suggests that FM and visual speech cues provide different speech information. If such information were independent, then it might be expected that the effect would be completely additive such that the score for the AM+FM AV condition (in noise) would be equal to the percent correct score for the AO AM condition (in noise) + the FM effect (AO AM+FM minus AO AM) + the visual enhancement effect (AV AM minus AO AM). This was in fact the case; the score for the AM+FM AV condition was 61.6% was approximately the same as 11% (AM) + 24% (AO AM+FM minus AO AM) + 27% (AV AM minus AO AM).

4. Discussion

Previous research has found that AM cues (even extracted from a limited number of speech frequency bands) can support correct speech recognition (with levels of more than 80% correct being achieved, e.g., Shannon et al, 1995). Recently, Zeng et al (2005) have demonstrated that additional fine-structure FM information is needed to maintain reasonable levels of speech identification when there is competing background noise.

The Zeng et al study was conducted only with auditory signals (the listeners did not see the talker speaking). The present study aimed to replicate the 8 sub-band Auditory-Only condition of Zeng et al and to further investigate whether there would still be an FM advantage if the listener had access to visual speech information. The results support those of Zeng and colleagues in demonstrating an FM advantage for speech recognition with auditory only presentation, an advantage that was greater with background noise conditions. The findings also show that the FM advantage still occurs when the listener is able to see the talkers face. What is more, the size of this FM effect was similar in the auditory only and Auditory-visual conditions suggesting that the information provided by FM and visual speech is different. The current results support the conclusion of Nie, Zeng and colleagues that the design of cochlear implant speech processors should aim to encode both AM and FM information even if visual speech cues are present.

5. Acknowledgements

The authors would like to thank Dr. Fang-Gang Zeng for use of the FAME signal processing algorithm. They also thank Darren Cullinane for help producing test stimuli and Dr Jeesun Kim for her comments on the manuscript. The second author also wished to acknowledge the support of the Australian Research Council.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

6. References

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, *35*, 116-124.

Nie, K., Stickney, G., & Zeng, F-G. (2005). Encoding Frequency Modulation to Improve Cochlea Implant Performance in Noise. *IEEE Transactions on Biomedical Engineering*, *52*, 64-73.

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. *Perception and Psychophysics*, *48*, 313-325.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. & 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.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America. 26*, 212-215.

Zeng, F-G, Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., & Cao, K. (2005). Speech recognition with Amplitude and Frequency Modulations. *PNAS*, *102*, 2293-2298.

Zeng, F-G. (2004). Trends in Cochlear Implants. *Trends in amplification*, *8*, 1-34.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review