Recognizing spoken vowels in multi-talker babble: Spectral and visual speech cues

Chris Davis and Jeesun Kim
MARCS Auditory Laboratories, University of Western Sydney, Australia

chris.davis@uws.edu.au; j.kim@uws.edu.au

Abstract

It has been proposed that both spectral and visual speech cues assist in segregating a talker from noise. To test how these cues interact, the experiment examined vowel identification (in hVd context) when presented in multi-talker babble. The availability of spectral cues was manipulated by filtering the signal into (1) 8 frequency amplitude-envelope bands or (2) the same bands with additional spectral cues. The availability of visual speech cues was manipulated by using auditory-only (AO) and auditory-visual (AV) presentations. It was found that the intelligibility benefit when spectral and visual speech cues were combined appeared to be less than that produced by adding the benefits for each cue type when tested separately. This pattern suggests that both cues provide similar information.

1 Introduction

Zeng, Stickney, Nie and colleagues have, over a number of papers, (e.g., [1], [2], [3]) demonstrated that the addition of frequency modulation (FM) to speech that was filtered so as to only provide amplitude envelope (AM) information in a limited number of frequency bands, assists listeners in segregating the utterances of a target speaker from those of others (i.e, in babble noise). It has been proposed that this FM benefit occurs because the additional acoustic FM information provides about formant trajectories and F0 allows the speech of the target talker to be better grouped and therefore tracked over time. In essence, what has been proposed is an auditory scene analysis account in which speech from a target talker can be more effectively parsed from background distractors.

A parallel scene analysis account has been proposed to account for the benefit in identifying the target speech when the target talker can be seen. The usefulness of visual speech information for source separation has been demonstrated in behavioral experiments by [4] and more recently by [5]. Furthermore, [6] and [7] have demonstrated that visual speech can provide a useful constraint for blind source separation.

The aim of current experiment was to examine what would happen in the case where FM and visual speech cues were both provided to a listener whose task was to identify speech in multi-talker babble noise. It seemed to us that seeing a target speaker would provide such a potent cue to segregating the talker that any additional FM cues to segregation may be redundant. If this was the case and the FM benefit was purely a product of talker segregation, then there should be little FM benefit shown when the listener can see the target talker. If, however, the FM cue provides additional information about the target speech itself, there may still be a sizeable FM benefit observed even though visual speech is provided.

2 Methods

2.1 Participants

Thirty-one undergraduate university students from the University of Western Sydney participated in the experiment. All participants were native speakers of English, 18 years of age or over and had self-reported normal or corrected-to-normal vision and none reported a history of hearing loss.

2.2 Stimuli

The stimuli consisted of 16 hVd syllables (in which V = /iː, I, æI, e, æ, ɔ, a, aː, oː, əu, ʊ, u, ɑe, oI, æɔ, ɜː/) spoken by a male native Australian English speaker and recorded in a double walled, sound attenuated room. The speaker was recorded against a uniform grey background, facing the camera and the recording showed the head and shoulders. The video was digitized at 29.9 frames per second with a resolution of 352 x 240 pixels. The audio component of the video was sampled at 44,000 Hz. A commercially available multi-talker (three female talkers and one male) babble track (Auditec, St. Louis, MO) was used as competing noise stimuli.

2.3 Signal processing

Audio signal processing was performed using the frequency amplitude modulation encoding (FAME) processing algorithm [1]. The following presents a brief description of this algorithm (see [2]). The wide-band signal was separated into 8 sub-bands by band-pass filters. For each sub-band the AM and FM signals were extracted (by Hilbert transform). The slow varying AM envelope was obtained by full wave rectification and low-pass filtered at 500 Hz. The slow varying FM signal was obtained by the removal of each band's centre frequency (by phase-orthogonal demodulators), a

subsequent low-pass filter (with a cutoff frequency of 500 Hz) and rate (with a cutoff value of 400 Hz). The delay between the AM and FM signals was adjusted and the AM and FM signals were each combined into their respective sub-bands. These signals were further bandpassed filtered to remove frequency components outside the original analysis filter's bandwidth. The band-passed signals were then summed to form the synthesized AM+FM signal. In the noise condition, the audio portion of both the target hVd syllable and competing 'babble' noise were mixed at -5dB signal to noise ratio (SNR) and subjected to FAME processing (i.e., FAME processing was applied after the target and masker were mixed). 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, the processed speech signal was dubbed onto the video of the talker.

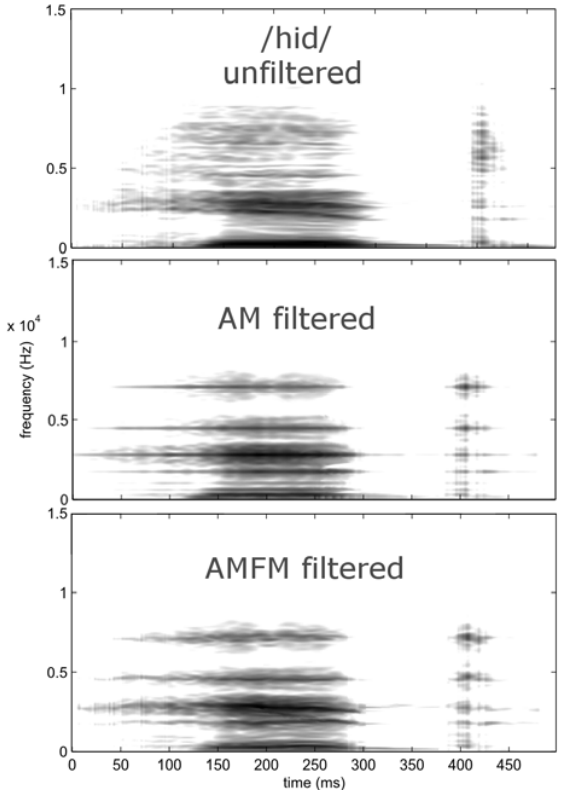

Figure 1. Spectrograms of the unprocessed word /hid/ (upper panel), the 8 band AM filtered version (middle) and the 8 band AM+FM filtered version (bottom). Note how the formants differ across the conditions.

Figure 1 shows a set of spectrograms that illustrate the characteristics of the AM and AMFM filter properties. As can be seen in the figure (middle panel), formant information is degraded by the AM filter. The addition of FM information has the effect of providing some formant spacing and trajectory information.

Figure 2 provides an illustrative example that shows that the power of the filtered stimuli remained relatively unaffected by the filtering operation but that F0 is affected. As can be seen, F0 can be tracked in the unfiltered stimulus (top panel) but tracking is compromised by the AM filter (middle) and less affected by the AMFM filter.

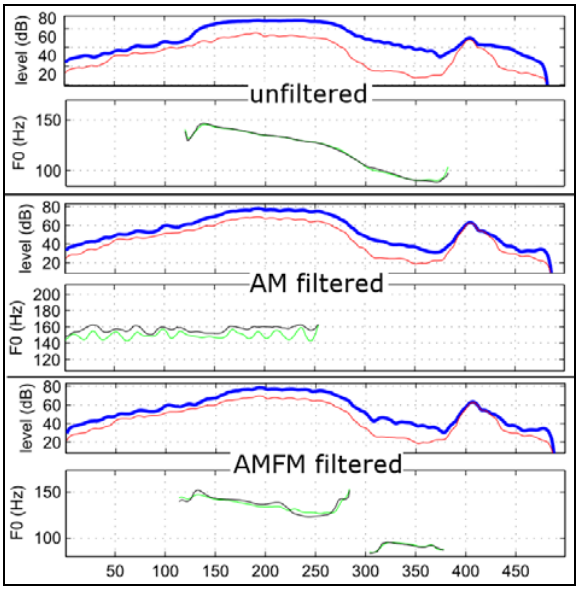

Figure 2. Shown pairs of power and F0 plots for the unfiltered stimulus /hid/ (top panels); the AM filtered version (middle) and the AMFM version (lower panels). In the upper panel of each pair, the thick line represents the total power, the thin line shows power in frequencies >3kHz.

2.3 Procedure

Participants were tested individually in a sound attenuated booth. Auditory stimuli were presented through Sennheiser HD580 headphones. The video clips (compressed to MPEG 2) were played back using the DMDX software ([8]) on a 21 inch monitor. The stimuli were presented under 4 conditions (each in noise): AO AM, AO AM+FM, AV AM and AV AM+FM. Four versions of the item list were prepared such that no item was repeated in any version but each version contained all conditions (i.e., syllables were fully rotated over conditions). A participant was allocated to two versions such that they were presented with the same item in the AM and AM+FM conditions. Note that items in the AO condition were not presented in the AV condition. This procedure was used to

minimize learning effects from the AV to the AO presentations. The AO and AV conditions were presented in blocks but presentation of items within each block was randomized, as well as the presentation of the condition blocks themselves. An open-response format was used.

3 Results

The mean percent correct vowel identification results are shown in Figure 3.

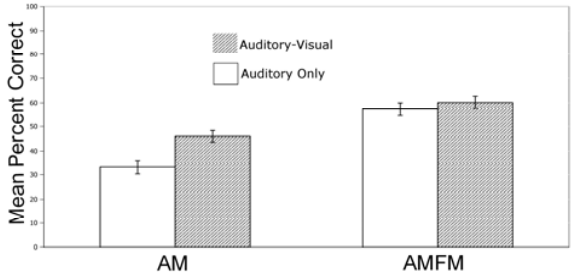

Figure 3. Mean percent correct vowel recognition scores for each condition (AM = AM cue; AMFM = AM + FM cue; AO = Auditory Only; AV = Audio-Visual Speech).

There was an overall visual speech facilitation effect; with more vowels being correctly identified in the AV condition (53.1%) compared to the AO (45.3%) condition, F1 (29) = 7.69, $p < 0.05$. There was also an FM effect: more vowels were correctly identified in the AMFM (58.7%) condition than in the AM (39.6%) condition, F1 (29) = 100.6, $p < 0.05$. There was also an interaction between the visual speech and FM facilitation effects, $F(1,29) = 5.26$, $p < 0.05$.

To determine the relative contribution of phonetic features to pattern of date, the identification data was re-scored in terms of features. The features used are shown in Table 1.

Table 1. Feature classification for the vowels

Label	Features		
Vowel Height	Closed	Mid	Open
Place	Front	Centre	Back
Lip Rounding	Rounded	Unrounded	

The score for each feature was obtained from AO and AV confusion matrices by assessing each response according to the stimulus feature classifications listed in Table 1.

The percent correct scores for each feature as a function of presentation condition and filter type are shown in Figure 4. The percent correct scores were analyzed in a series of ANOVAs. For the lip-rounding feature, there was a significant AV and FM effect (15.8%), $p < 0.05$. For the vowel height feature, the AV effect (1.1%) was not significant, $F < 1$ but FM effect was significant, $p < 0.05$. For the place feature there was an AV effect, $p < 0.05$ and an FM effect, $p < 0.05$. Further analyses showed that there was a larger AV effect if the feature was a front vowel compared to a centre or back vowel, $p = 0.05$; the size of the FM facilitation effect was the same for the front and centre/back vowels.

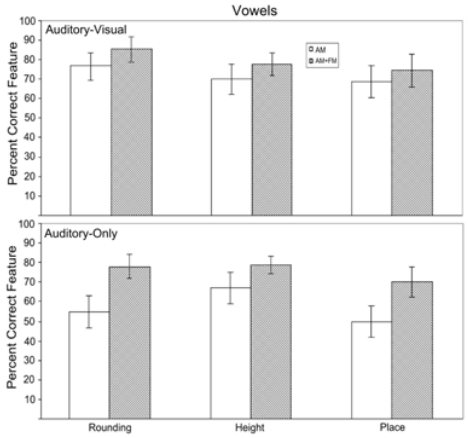

Figure 4. Mean percent correct vowel recognition scores in terms of features for the various presentation conditions.

The results can be considered either in terms of how the visual speech or the FM facilitation effects varied over the presentation and filter conditions. In general, it appears that there was a smaller FM effect for AV compared to AO presentation. It appears that the visual speech facilitation effect was smaller when scored in terms of vowel height compared to a feature like lip rounding.

4 Discussion

When the speech of a target talker is masked by the speech of others, the intelligibility of AO frequency band limited speech (i.e., AM filtered speech - often used as a model of how speech is coded by cochlear implants) is improved by the addition of frequency modulation. In a comparable effect, intelligibility in noise can be increased when the target talker can be seen. The current study investigated what would happen to the intelligibility of frequency band limited speech in noise when both FM cues and visual speech are available.

It has been proposed that the FM advantage in intelligibility is likely driven by its contribution to assisting in scene analysis (via better resolution of target

F0) and by its contribution to source extraction per se (possibly formants) [3]. The facilitation in intelligibility provided by visual speech probably also entails a twin contribution: assisting with scene analysis and a speech-reading component. If the addition of both FM and visual speech cues produced a boost in intelligibility equal to the size of the sum of each separate effect it would indicate that the FM and visual speech cues provide non-redundant information (e.g., that speech-reading provides different information than the source information provided by the FM cue).

The results showed that the boost in intelligibility from both FM and visual speech cues was less than the sum of the separate effects. That is, whereas AV presentation produced a robust facilitation effect for AM filtered speech, this was significantly reduced for the AMFM filtered speech (see Figure 3). This reduction in the size of visual speech facilitation seems unlikely to be the results of a ceiling effect since performance in the AV AMFM condition was only at 60% correct.

The above description of results was presented in terms of a reduction in the size of visual speech facilitation; however alternatively, the results could have been couched in terms of a reduction in FM facilitation. That is, because the AV AMFM presents both types of cues, it is not possible to determine the precise contribution of each to the percent correct identification score. One thing is clear however, the facilitation effect for the combined cues was less than what might have been expected from the size of the visual speech effect in the AM condition and the size of the FM effect in the AO condition.

This pattern of under-additivity suggests that the visual speech and FM cues might be providing similar information. This finding is consistent with the interpretation that both cue types allow a perceiver to perform more effective auditory scene analysis and thus do better in parsing the target talker from the background talkers. One difficultly with taking this as a general formula regarding the way that FM and visual speech cues will interact is that recently [9] has demonstrated that the FM and visual speech cues can combine in an additive fashion for different types of speech (e.g., for consonants). It is possible that additive FM and visual speech effects occur when the source extraction benefit provided by FM cues affords complementary information to that supplied by speech-reading.

5 Acknowledgements

The authors wish to acknowledge the support of the Australian Research Council (ARC) grant DP0666857 and the ARC and NHMRC grant TS0669874.

6 References

[1] 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.

[2] 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. (2005) PNAS, 102, 2293-2298.

[3] Stickney, G. S., Nie, K., and Zeng, F. G. (2005). "Contribution of frequency modulation to speech recognition in noise," J. Acoust. Soc. Am. 118, 2412-2420.

[4] Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature, 381, 66-68.

[5] Helfer, K. S., & Freyman, R.L. (2005). The role of visual speech cues in reducing energetic and informational masking. J. Acoust. Soc. Am. 117, 842–849.

[6] Sodoyer, D. Girin, .L. Jutten C, & Schwartz, J.-L. (2004). Developing an audio-visual speech source separation algorithm. Speech Commun. 44, 113–125.

[7] Rivet, B., Girin, L., and Jutten, C. (2007). Visual voice activity detection as a help for speech source separation from convolutive mixtures. Speech. Commun. 49, 667-677.

[8] Forster, K. I., and Forster, J. C. (2003). "DMDX: A windows display program with millisecond accuracy," Behav. Res. Meth. Instr. Comp. 35, 116-124.

[9] Kim, J., Davis, C., & Groot, C. (under revision). Speech identification in noise: Contribution of temporal, spectral and visual speech cues