Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Modification of prosodic cues when an interlocutor cannot be seen: The effect of visual feedback on acoustic prosody production

Erin Cvejic, Jeesun Kim and Chris Davis

MARCS Auditory Laboratories, University of Western Sydney, Milperra, Australia

PACS: 43.70.Fq; 43.70.Mn

ABSTRACT

Speakers alter the way they produce speech according to the communicative situation. Changes are made to enhance the efficiency of information transmission. For instance, when in noisy environments, people speak loudly and produce more energy in higher frequencies (the Lombard effect). This study investigated whether a change in the visual conditions associated with communication would also lead to modification in speech production. More specifically, it examined if auditory prosody would be affected by whether the speaker could see the interlocutor or not. In the experiment, two types of prosodic contrasts were included. The first was 'prosodic focus' used by speakers to enhance the perceptual salience of an item. The second was 'prosodic phrasing' which refers to the phrasing of a sentence as a question without using an interrogative pronoun. Four speakers were recorded while completing a dialog exchange task in which the interlocutor could or could not be seen. The results showed that the corner-most vowels recorded in narrow focus and echoic question contexts were produced over longer durations and with a greater vowel space (reflected by greater vowel triangle area and vowel triangle dispersions) relative to broad focused renditions across both interaction conditions. With the exception of intensity, no other acoustic or spectral properties appeared to be enhanced at the phonemic level when the interlocutor was not visible to the speaker. This may be due to prosody affecting the utterance at more global levels (e.g., word and utterance levels), rather than at the localized vowel level. That is, modifications may be seen between interactive conditions in terms of pitch contours, pre-focal shortening and intensity profiles when examined across the whole utterance.

I. INTRODUCTION

It is known that speakers make situation- and audience-dependant changes to the speech signal that they produce in order to increase the likelihood of the audience clearly understanding the intended message. That is, speakers adjust the way that they speak depending on who (or what) they are talking to, and the environment in which they interact. For example, when conversing with infants, adults engage in so-called "infant-directed speech", characterised by increased pitch, greater perceptually rated affect and hyperarticulation of vowels (i.e., expansion of the F1-F2 vowel space) compared to the speech directed towards another adult [1, 2]. Similarly, in the presence of noise (i.e., Lombard speech), speakers increase loudness and pitch, and decrease their speaking rate relative to production of speech in quiet situations [3]. Of interest in this study is whether a change in the *visual* conditions associated with communication would also lead to modifications in speech production. More specifically, it examined whether the production of auditory prosody would be affected by whether the speaker could see the interlocutor or not.

Prosody is a broad term used to describe variations in the auditory speech signal corresponding to the perception of pitch, loudness and duration. Of its many functions, prosody can indicate general speaker characteristics (such as gender, age, physiological and emotional states), assist in the segmentation of an incoming speech signal into meaningful units allowing for understanding, and convey information extending beyond that provided by sentence syntax, grammar and the symbolic content of speech sounds alone [4]. Put simply, the modification of suprasegmental acoustic cues can alter the linguistic message, without manipulating the syntactic content of an utterance. Two such contrasts of interest in this study are prosodic focus and prosodic phrasing.

Prosodic focus describes the situation where a word is made perceptually more salient than other words within a sentence, and is used to emphasize importance or to disambiguate a particular constituent within an utterance. The focused item within such an utterance is said to be "narrowly focused", as the point of informational focus has narrowed down to that particular item [5]. Narrow focus contrasts with "broad" focused renditions, where there is no explicit point of informational focus. The prosodic type we refer to as prosodic phrasing refers to acoustic modifications used to achieve different sentence types, such as statements and questions. By mimicking the syntactic content of a declarative statement but altering suprasegmental parameters, an "echoic" question can be phrased without the use of an interrogative pronoun [6]. That is, although echoic questions contain the same syntactic content as a declarative statement, a level of uncertainty can be implied through the manipulation of acoustic cues.

The acoustic properties associated with prosodic focus and prosodic phrasing have been intensively studied and well described. In general, a narrowly focused word (relative to the same word produced in a broad focused rendition) tends to be articulated with a higher fundamental frequency (F0), have a greater intensity and consist of longer syllable durations [7]. Narrowly focused vowels also appear to be produced with greater first formant values than the same vowel produced in broad focused contexts [8].

Different sentence phrasings typically vary in the following ways: statements can be characterised as having a steadily falling F0 contour ending with a sharp drop (often signalling finality), whereas the opposite pattern is observed for echoic questions (i.e., gradually rising F0 throughout the time course of the utterance, with a final sharp rise in pitch, indicating a response may be required from an interlocutor) [6]. Statements also tend to have shorter syllable durations and steeper falls in final intensity compared to the same sentences uttered as echoic questions. In addition to affecting the utterance at a global level, echoic questions can also be deemed to have a narrow point of informational focus (i.e., one particular constituent that is questioned within an utterance) [9]. A questioned word differs from broad focused renditions of the same word in terms of pitch contour and is typically produced over an increased duration (see Figure 1 for a comparison of time-normalised pitch contours for the same sentence produced with broad focus, narrow focus and echoic question renditions).

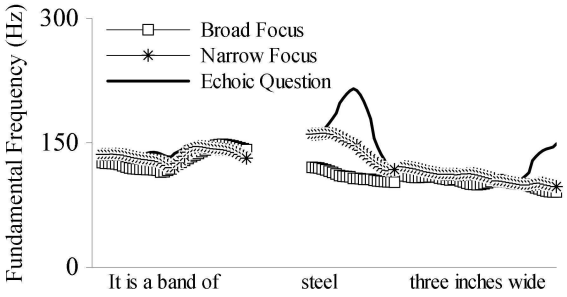

Figure 1. Time-normalised F0 patterns for three renditions (broad focus; narrow focus; echoic question) of the sentence "It is a band of steel three inches wide" uttered by a male speaker. The word "steel" is the critical word that receives narrow focus or question intonation. The prosodic speech condition affects the utterance at both the local level of the critical word, as well as at the global level of the utterance. (Source: Cvejic et al., 2010 [10])

Speakers are receptive to the linguistic needs of their audience, and are able to tailor the produced speech signal accordingly (whether this being due to the interactive environment or the nature of the intended message). However, the interlocutor also plays a role in eliciting these modifications by providing the speaker with a range of feedback cues. Such cues can be used by the speaker to make adjustments to the signal in order to increase the chance that the intended message is received by the interlocutor [11]. While these cues may be acoustic (e.g., explicitly asking for clarification, or back-channelling to indicate understanding), they can also be visual. That is, when interacting with people, a wealth of visual cues are available from a listener's face and from bodily gestures (moreover, the visual cues from the face of a speaker are also a rich source of information for an interlocutor, and are useful for both segmental [12, 13] and suprasegmental tasks [10, 12]). In this study, we were interested in examining the effect that being able to see the interlocutor has on the production of acoustic prosody. That is, are the production of acoustic cues different when a speaker can see the person they are talking to compared to when they cannot?

II. METHOD

A. Materials

The materials consisted of 30 non-expressive, phonetically balanced sentences drawn from the IEEE Harvard Sentence list [14], describing mundane events with minimal emotive content. Each sentence was recorded in three prosodic conditions: as a *broad focused* statement, a *narrow focused* statement, and as an *echoic question*.

To elicit the conditions in this study, a dialogue exchange task was used [10, 15] requiring the speaker to interact with an interlocutor, and either repeat what they heard the interlocutor say (broad focused statement) make a correction to an error made by the interlocutor (narrow focused statement, Example 1), or question an emphasized item within the sentence produced by the interlocutor (echoic question, Example 2). An example of this dialogue is given below:

Example 1.
I: It is a band of [rubber]$_{\text{ERROR}}$ three inches wide.
S: It is a band of [steel]$_{\text{CORR.}}$ three inches wide.

Example 2.
I: It is a band of [steel]$_{\text{EMPH.}}$ three inches wide.
S: It is a band of [*steel*]$_{\text{QUEST.}}$ three inches wide?

B. Apparatus

Auditory data was captured using a Behringer C-2 condenser microphone placed approximately 30cm below the speaker's mouth, held in position with a boom-arm microphone stand. The acoustic signal was sampled at 44.1 kHz digitized mono.

C. Procedure

Each speaker was recorded individually while seated in an adjustable dentist's chair within a double-walled, sound insulated booth. In this study, two interaction conditions were used; face-to-face (FTF) and audio-only (AO) interactions.

In the FTF condition, participants were instructed to direct their speech towards an interlocutor, who was located approximately 2.5 meters in front of them (Figure 2) while engaging in the dialogue exchange task previously outlined. Participants were instructed to speak as naturally as possible. This condition allowed both the speaker and interlocutor to see and hear each other. The interlocutor was not instructed to provide any additional visual cues to the interlocutor, nor was the speaker instructed to pay special attention to the visible cues available from the interlocutor.

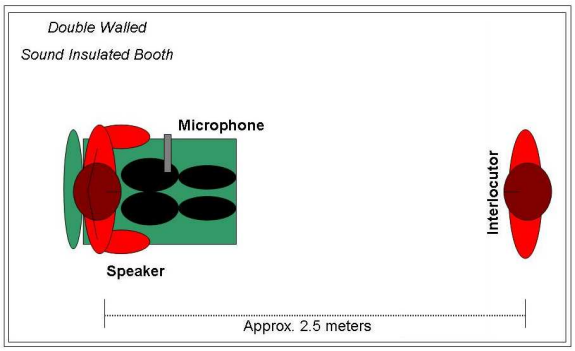

Figure 2. The experimental setup used in the FTF interaction condition. The speaker and interlocutor engaged in the dialogue exchange task while visible to each other.

In the AO interaction condition, participants completed the dialog exchange task over a microphone and head-phone setup shown in Figure 3. Behringer C-2 condenser microphones were inputted into the left (speaker) and right (interlocutor) audio channels of an Edirol UA-25 USB audio capture card. The left channel was then outputted to the interlocutor, and the right channel outputted to the speaker, though Senheiser HD650 stereo headphones. The input sensitivity was adjusted so that the speaker and interlocutor could hear each other at a perceived comfortable, conversational volume. Participants remained seated in the dentists chair as for the FTF condition; however the interlocutor was located outside of the room (Figure 4). This condition allowed the speaker and interlocutor to interact with each other, but this interaction was restricted to only auditory communication (i.e., visual feedback was no longer available from the speaker).

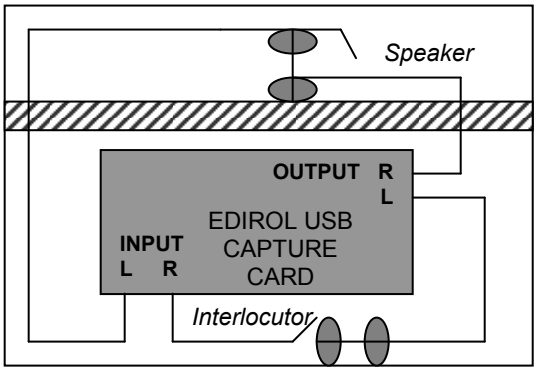

Figure 3. Audio setup used in the audio-only interactive condition, allowing for the speaker and interlocutor to converse over microphone and headphone setup without being visible to each other.

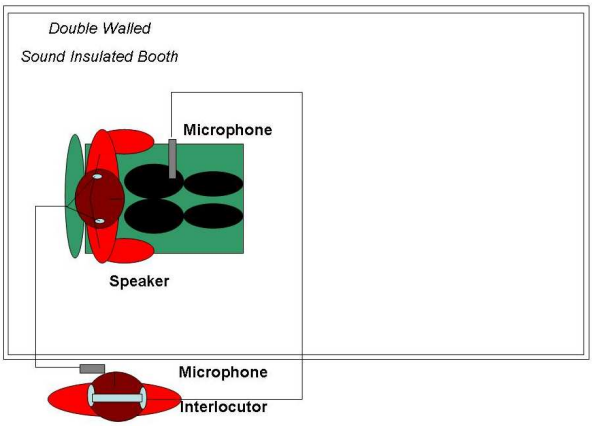

Figure 4. The experimental setup used in the AO interaction condition. The speaker engaged in the dialogue exchange task with the interlocutor over a microphone and head-phone setup, allowing them to hear each other while not being visible to each other.

In each interaction condition, two repetitions of each sentence were recorded in each of the three prosodic speech conditions. In total, 360 sentences (30 sentences x 3 prosodic conditions x 2 repetitions x 2 interaction conditions) were recorded for each speaker.

D. Participants

Four male speakers (M_{Age}= 22.5 years) participated in the data capture sessions. All were native speakers of Standard Australian English, with similar levels of education. All speakers had self-reported normal hearing and vision, with no known communicative dysfunction.

E. Data processing

Captured auditory data was subjected to semi-automatic forced phonemic alignment using the MARY-TTS engine [16], before manually correcting the identified phoneme boundaries in Praat [17]. For the purpose of this study, we were concerned with the fine-grained acoustic-phonetic characteristics associated with the production of prosody across different interactive conditions, and investigated these by examining changes in the vowel space.

In order to measure the vowel space of each speaker, a subset of the three most corner vowels from the stimuli /a,i.o/ were selected from within the critical words. The sentences, and the associated critical vowels, are shown in Table 1.

Table 1. Subset of the 10 sentences containing the target vowels from which the vowel space measurements were calculated. The critical word is within brackets.

/a/:
It was hidden from sight by a [mass] of leaves and shrubs.
The weight of the [package] was seen on the high scale.
Hold the [hammer] near the end to drive the nail.
/i/:
It is a band of [steel] three inches wide.
The lobes of her ears were [pierced] to hold rings.
A [pink] shell was found on the sandy beach.
This is a grand [season] for hikes on the road.
/o/:
Clams are round, [small], soft and tasty.
A [small] creek cut across the field.
The set of china hit the [floor] with a crash.

Fundamental frequency, mean relative intensity and durational properties of the critical vowels were extracted using custom-designed scripts in Praat [17]. Steady state formant properties (F1 and F2) were extracted using the procedure outlined in [18]. The acoustic data was initially down sampled to 10 kHz, before calculating formant frequencies by applying a 25ms sliding window (with steps of 1ms) to the signal, with the steady state value being determined by averaging 40% of the formant estimates between 40 and 80 percent of the total vowel duration.

The obtained formant values were then converted to the perceptually motivated Mel scale [19], using (1):

$$M = (1000/\log 2)\times(\log((F/1000)+1)) \quad \textbf{(1)}$$

where M and F are frequency values expressed as Mels and Hertz, respectively [20].

In order to calculate the size of the vowel triangle, the Euclidean distance between each vowel category centre was obtained. The area of the generated vowel triangle can then be calculated using (2):

$$\text{VTArea (Mels}^2) = \sqrt{(S\times(S-A_L)\times(S-B_L)\times(S-C_L))} \quad \textbf{(2)}$$

where $S = ((A_L+B_L+C_L)/2)$, and A_L, B_L, and C_L are the Euclidean distances in Mels between vowel category centres /a/ to /i/, /i/ to /o/ and /o/ to /a/, respectively.

To calculate the vowel space dispersion, the Euclidean distance was calculated between the vowel space midpoint, and each vowel token, indicating the overall expansion (or compaction) of the vowel space. The mean of these distances is reported. Within category dispersion was calculated in a similar way, by determining the Euclidean distance between the midpoint for each vowel category, and each measured

token within that category, with mean obtained for these values. This measure provides an indication of individual vowel category dispersion, which may indicate consistency or variability of individual vowel productions across repetitions within each prosodic category and interactive condition [20]. All distance calculations were carried out in Matlab (The Mathworks, Inc.).

In the following analyses, we first compared the acoustic and spectral properties across prosodic conditions in the FTF condition, then examined these properties for the condition where interaction occurred within an AO context. Finally, we examined the differences between the two interactive conditions.

III. RESULTS

A. Acoustic analysis in FTF condition

Table 2 outlines the fundamental frequency, relative intensity and duration of the vowels /a, i, o/ as a function of the prosodic context collapsed across speakers. Even at the phonemic level, differences across the prosodic conditions can be observed (most notably for durational properties).

Table 2. Acoustic properties of the corner-most vowels /a, i, o/ as a function of prosodic context, recorded in the FTF interactive condition. Values collapsed across speakers.

Vowel	Prosodic Condition		
	Broad Focus	Narrow Focus	Echoic Question
Fundamental Frequency (Hz)			
/a/	113.00	117.13	109.92
/i/	120.00	125.78	111.28
/o/	109.21	116.00	115.17
Relative Intensity (dB)			
/a/	62.61	64.69	62.32
/i/	63.50	65.81	63.70
/o/	63.59	65.80	64.35
Duration (ms)			
/a/	95.32	123.67	116.84
/i/	97.77	135.97	132.44
/o/	137.96	214.80	182.72

To determine whether there were significant differences on these acoustic properties, these data were collapsed across vowel categories, and subjected to a series of planed comparison paired samples *t*-tests. Overall, the fundamental frequency of vowels was significantly higher in narrow focus than broad focus renditions, $t(79) = 7.52$, $p < .001$, however no significant difference was observed between broad focus and echoic questions for the fundamental frequency of steady state vowels, $t(79) < 1$, $p > .05$. Similarly, the mean relative intensity of narrow focused vowels was greater than that for broad focused vowels, $t(79) = 10.89$, $p < .001$, whereas no different was observed between broad focus and echoic question renditions, $t(79) = 1.31$, $p = .194$. In terms of duration, vowel lengths in both narrow focus ($t(79) = 11.13$, $p < .001$) and echoic question contexts ($t(79) = 8.94$, $p < .001$) were significantly larger than the same vowels recorded within a broad focused context.

It should be noted that the above observations pertain to measured differences at the vowel level, and as such, acoustic differences that are often used to characterise prosodic conditions (e.g., fundamental frequency modulation particularly for echoic question renditions) were not found. Such differences are however likely to be seen when examining utterances at a broader constituent level (i.e., at the word or sentence level).

Table 3 outlines the area of the vowel triangle space, vowel space dispersion, within category dispersion, and F1 and F2 ranges for vowels recorded in the FTF interaction condition for each individual speaker. From this and Figure 5, it can be seen that the vowel triangles across speakers generally expand as a function of the prosodic conditions, larger in narrow focus and echoic question renditions compared to the same vowels produced within broad focused utterances.

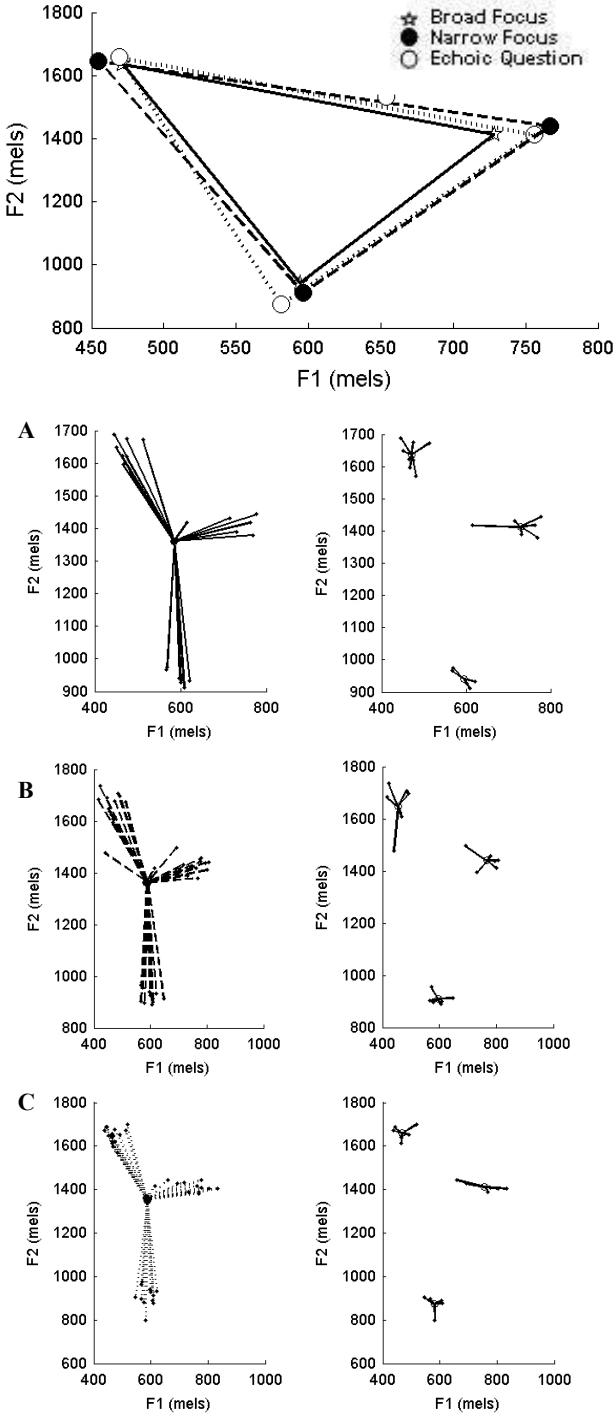

Figure 5. Vowel triangle properties for broad focus, narrow focus and echoic questions in the FTF interactive condition for Speaker 1. The top panel shows vowel triangle expansion between the conditions, the left column indicate displacement from the F1-F2 vowel space midpoint, and the right column shows within-category displacement from vowel category midpoints for (A) broad focus, (B) narrow focus and (C) echoic question renditions

Table 3. Spectral properties (divided by speaker and prosodic condition) in the FTF interaction condition.

Speaker	Prosody	Spectral Measure				
		VT Area ($Mels^2$)	VT Dispersion (Mels)	Within Category Clustering (Mels)	F1 Range (Mels)	F2 Range (Mels)
1	Broad	75573.12	292.47	39.16	330.80	777.80
	Narrow	100326.92	322.30	49.61	389.95	846.61
	Echoic	98693.31	330.74	39.59	396.01	898.75
2	Broad	146520.28	375.82	65.12	417.23	1109.58
	Narrow	155996.32	411.10	143.09	532.86	1146.72
	Echoic	179797.62	428.14	58.30	440.98	1141.74
3	Broad	55790.97	261.77	78.81	329.37	876.60
	Narrow	49414.21	306.11	113.28	371.23	1063.51
	Echoic	55997.80	287.45	158.09	460.17	996.47
4	Broad	94084.75	333.27	70.15	336.24	952.34
	Narrow	144561.43	397.17	65.65	476.99	1052.45
	Echoic	141307.99	389.33	56.56	432.44	1015.11

When producing the vowels within these contexts, it appears that speakers make an increased effort to make the vowels more distinct compared to when produced in a broad focused context. This finding is reflected in the VT dispersion measure, with greater vowel distances from the midpoint of the F1-F2 space when a word in narrow focusing or questioned within an utterance.

Within category clustering also appears to expand in narrow focus and echoic question conditions; however this appears to be highly variable across speakers. Finally, the range of F1 and F2 covered also appears to increase in narrow focus and echoic question renditions, compared to broad focus productions. A greater vowel space is covered in these conditions, which may be part of the speakers' strategy to ensure that their intended message is being clearly understood by the interlocutor.

B. Acoustic analysis of AO interaction condition

Table 4 outlines the acoustic properties of the corner vowels collapsed across speakers, as a function of the prosodic speech condition in the AO interactions. As with the FTF interaction condition, there were noticeable differences even at the phonemic level between the prosodic conditions, most noticeably in terms of vowel length.

Table 4. Acoustic properties of the corner vowels as a function of prosodic context, recorded in the AO interactive condition, collapsed across speakers.

Vowel	Condition		
	Broad Focus	Narrow Focus	Echoic Question
Fundamental Frequency (Hz)			
/a/	115.05	120.71	107.30
/i/	115.94	135.31	114.56
/o/	109.08	109.08	124.88
Relative Intensity (dB)			
/a/	61.76	64.40	61.82
/i/	62.88	65.80	63.00
/o/	62.23	66.12	63.21
Duration (ms)			
/a/	91.10	125.16	118.60
/i/	99.30	137.85	129.91
/o/	140.00	216.21	195.82

Planned-comparison paired samples *t*-tests (collapsed across vowel categories) revealed that fundamental frequency was significantly greater in narrowly focused than broad focused renditions, $t(79) = 2.50$, $p = .014$. The difference between broad focus and echoic question renditions was not significant. Similarly, mean relative intensity was significantly greater for narrow focus than broad focus renditions, $t(79) = 7.38$, $p < .001$, but not for echoic questions relative to broad focus items. As with the FTF interaction condition, significant durational differences were found between the narrow and broad focused vowel lengths ($t(79) = 9.88$, $p < .001$) and between echoic question and broad focus items ($t(79) = 7.90$, $p < .001$). These results are consistent with the findings of the face to face interactive condition.

Table 5 shows the spectral properties of the corner vowels from the stimulus set in the AO interaction condition. Once again, vowel space across different speakers shows expansion between broad focus and both narrow focus and echoic question utterances (see Figures 6 and 7). However, the overall area of the vowel space, and degree of expansion, appears to be highly variable across speakers. The dispersion of the vowel space generally increases in narrowly focused and echoic question conditions compared to broad focused renditions, with speakers producing these vowels with a greater distance from the vowel space midpoint. Within vowel category clustering is also highly variable across speakers and conditions, with no clear pattern of data (possibly due to the small number of observations made).

C. Comparison of interactive conditions

To compare the acoustic properties across the two interactive conditions, a series of planned comparison, paired-samples *t*-tests (with a Bonferroni adjusted alpha of .025) were conducted comparing fundamental frequency, relative intensity and duration between the two interactive conditions. For fundamental frequency, the only significant difference between the interactive conditions occurred for narrow focus, with F0 on average being lower in the AO interactive condition, $t(79) = 5.20$, $p < .001$, than in the FTF condition.

While no differences were observed between the broad focused intensity properties, speakers were louder in their vowel production in the AO than AV interaction condition for both narrow focus ($t(79) = 2.51$, $p = .014$) and echoic question prosodic conditions ($t(79) = 2.44$, $p = .017$). No durational differences were observed between the interaction conditions for any of the prosodic contrasts.

Table 5. Spectral properties (divided by speaker and prosodic condition) in the AO interaction condition.

Speaker	Prosody	Spectral Measure				
		VT Area (Mels^2)	VT Dispersion (Mels)	Within Category Clustering (Mels)	F1 Range (Mels)	F2 Range (Mels)
1	Broad	76886.87	295.16	46.14	334.64	802.92
	Narrow	92219.65	319.61	47.78	395.80	858.17
	Echoic	117332.23	342.84	42.03	398.69	865.99
2	Broad	154829.37	388.01	52.75	441.96	1069.80
	Narrow	189769.76	420.14	79.07	513.23	1160.49
	Echoic	187736.42	423.87	48.07	475.16	1124.75
3	Broad	41597.68	253.35	135.03	399.97	786.81
	Narrow	77854.98	333.89	89.54	362.33	1126.52
	Echoic	30493.23	250.83	140.88	389.39	897.41
4	Broad	87166.99	318.37	80.52	341.87	1047.73
	Narrow	129916.78	374.58	54.48	453.22	1008.73
	Echoic	126370.45	382.69	60.81	411.30	1033.93

To examine if there were any statistical differences between the interaction conditions in terms of vowel space, a 2×3 analysis of variance (ANOVA) was conducted, with interactive condition as a within-subjects factor, and prosodic condition as a between subjects factor. The pattern of vowel triangle areas across the different interactive conditions appears to be highly similar, with no significant main effects or interaction (F's < 1). This suggests that even when the speaker cannot see the interlocutor, the spectral properties of the speech signal remain relatively consistent. A difference was still observed for vowel space *within* the interactive conditions across the prosodic conditions, and these differences appear to be maintained, but not exaggerated, when speakers are engaging in tasks through auditory communication alone.

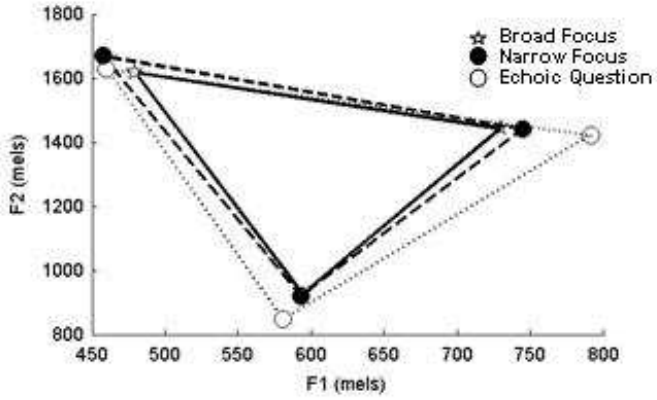

Figure 6. Vowel triangles in the AO interactive condition in broad focused, narrow focused and echoic question conditions for Speaker 1. The vowel triangle area was greater in the narrow focus and echoic question conditions compared to broad focused one.

IV. GENERAL DISCUSSION

The purpose of the current study was to investigate whether a change in the visual conditions associated with communication would lead to a modification in speech production, by examining whether the production of auditory prosody would be affected by the speaker being able to see the interlocutor or not. In general, both acoustic and fine-grained spectral properties of the corner vowels patterned in similar ways across both interactive conditions as a function of prosodic context. In terms of vowel space across speakers, a general

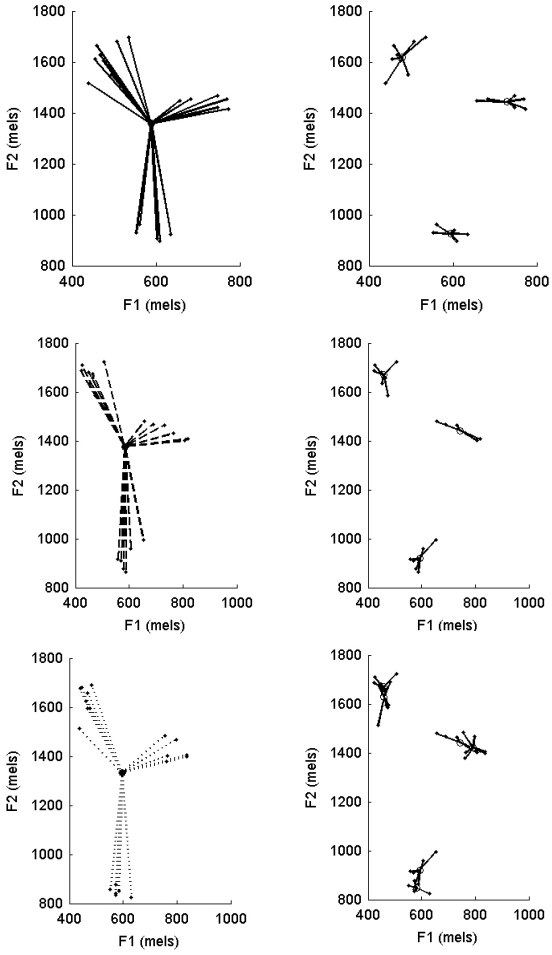

Figure 7. Vowel triangle displacement from the F1-F2 midpoint (left) and within-category dispersion from vowel category midpoints (right) in the AO interactive condition in (A) broad focused, (B) narrow focused and (C) echoic question conditions for Speaker 1.

expansion of the vowel triangle was observed across the prosodic conditions (i.e., vowels produced in narrow focus and echoic question contexts were more widely dispersed and covered a wider F1-F2 space relative to the same vowels within broad focused contexts). Similarly, vowel durations were elongated in narrow focus and echoic question contexts relative to broad focused renditions in both interactive conditions. However, the comparison of these properties across the

interaction conditions indicated no major differences between face to face and audio only interactions in the production of acoustic prosody.

Although no differences were found at the phonemic level, prosody impacts on an utterance at a more global utterance level. That is, there may be differences in terms of more global constituent and utterance features, such as pre-focal durational shortening, overall pitch contours and intensity profiles. Further acoustic analyses at a more global level (e.g., critical word, utterance) are required to obtain a better understanding of the role of visual feedback (if any) in the production of acoustic prosody. While no differences were found between the interactive conditions, considerable inter-speaker variation was observed across both acoustic and spectral properties. In this regard, the collection of more speaker data would be beneficial in determining just how much variation there is in the acoustic realisation of prosody and whether there are particular condition factors.

It is worth pointing out that in the AO interaction condition, participants wore headphones to interact with the interlocutor, which may have compromised the naturalness of speech production. That is, when wearing headphones, participants own acoustic feedback may have been reduced, requiring them to speak louder to perceive their own voice. Incorporating the use of headphones into the FTF interactive condition, and adding a self-feedback loop into the acoustic channel (e.g., speaker can hear themselves through the headphones in addition to the interlocutor) may provide better experimental control and result in more interesting findings.

ACKNOWLEDGEMENTS

The authors wish to thank Catherine Gasparini for her assistance with the recording procedure, and the four speakers for their time and patience. The first author acknowledges financial support from MARCS Auditory Laboratories. The second and third authors acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

REFERENCES

1 D. Burnham, C. Kitamura and U. Vollmer-Conna, “What’s new pussycat? On talking to babies and animals” *Science* **296**, 1435 (2002)
2 C. Lam and C. Kitamura, “Speech to infants with a simulated hearing loss” *J. Acoust. Soc. Am.* **125**, 2533 (2009)
3 J.C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers” *J. Acoust. Soc. Am.* **93**, 510-524 (1993)
4 S. Nooteboom, “The prosody of speech: melody an rhythm” in *The Handbook of Phonetic Science* ed. W.J. Hardcastle and J. Laver, (Blackwell, London, 1997) pp. 640-673
5 D. Bolinger, “Accent is predictable (if you’re a mind reader)” *Language* **48**, 633-644 (1972)
6 S.J. Eady and W.E. Cooper, “Speech intonation and focus in matched statements and questions” *J. Acoust. Soc. Am.* **80**, 402-415 (1986)
7 E. Krahmer and M. Swerts, “On the alleged existence of contrastive accents”, *Speech Commun.* **34**, 391-405 (2001)
8 W.V. Summers, “Effects of stress and final-consonant voicing on vowel production” *J. Acoust. Soc. Am.***82**, 847-863 (1987)
9 M.D. Pell, “Influence of emotion and focus location on prosody in matched statements and questions” *J. Acoust. Soc. Am.* **109**, 1668-1680 (2001)
10 E. Cvejic, J. Kim and C. Davis, “Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion” *Speech Commun.* **52**, 555-564 (2010)
11 S.E. Brennan and E. Hulteen, “Interaction and feedback in a spoken language system: A theoretical framework” *Knowledge-Based Systems* **8**, 143-151 (1995)
12 C.R. Lansing and G.W. McConkie, “Attention to facial regions in segmental and prosodic speech perception tasks” *J. Speech Lang. Hearing Res,* **42,** 526-539 (1999)
13 C. Davis and J. Kim, “Audio-visual speech perception off the top of the head” *Cognition* **100**, B21-B31 (2006)
14 IEEE Subcommittee on Subjective Measurements, “IEEE recommended practices for speech quality measurements” *IEEE Trans. Audio Electroacoust.***17**, 227-246 (1969)
15 E. Cvejic, J. Kim, C. Davis and G. Gibert, “Prosody for the eyes: Quantifying visual prosody using guided principal component analsysis” *INTERSPEECH 2010,* submitted (2010)
16 M. Schröder and J. Trouvain “The German text-to-speech synthesis system MARY: A tool for research, development and teaching” *Int. J. of Speech Tech.* **6,** 365-377 (2003)
17 P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer” *Glot Int.* **5,** 341-345 (2001)
18 K.G. Munhall, E.N. MacDonald, S.K. Byrne and I. Johnsrude, “Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate” *J. Acoust. Soc. Am.* **125**, 384-390 (2009)
19 G. Fant, *Speech sounds and features* (MIT Press, Cambridge, Massechusetts, 1973)
20 A.R. Bradlow, G.M. Toretta and D.B. Pisoni, “Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics” *Speech Commun,* **20**, 255-272 (1996)