INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The effect of spectral profile on the intelligibility of emotional speech in noise

Chris Davis[1], *Chee Seng Chong*[1], *Jeesun Kim*[1],

[1]The MARCS Institute, Western Sydney University, Australia

Chris.davis, l.chong, J.kim@westernsydney.edu.au

Abstract

The current study investigated why the intelligibility of expressive speech in noise varies as a function of the emotion expressed (e.g., happiness being more intelligible than sadness), even though the signal-to-noise ratio is the same. We tested the straightforward proposal that the expression of some emotions affect speech intelligibility by shifting spectral energy above the energy profile of the noise masker. This was done by determining how the spectral profile of speech is affected by different emotional expressions using three different expressive speech databases. We then examined if these changes were correlated with scores produced by an objective intelligibility metric. We found a relatively consistent shift in spectral energy for different emotions across the databases and a high correlation between the extent of these changes and the objective intelligibility scores. Moreover, the pattern of intelligibility scores is consistent with human perception studies (although there was considerable individual variation). We suggest that the intelligibility of emotion speech in noise is simply related to its audibility as conditioned by the effect that the expression of emotion has on its spectral profile.

Index Terms: speech recognition in noise, emotion, glimpses, masking

1. Introduction

This study examined why emotional expression affects speech intelligibility in noise, a result that has been reported in a number of recent studies e.g., [1; 2]. We considered two issues. (1). What acoustic correlate(s) of emotional expression best explains why speech expressing some emotions (e.g., happy) is more intelligible in noise than speech expressing other emotions (e.g., sad), even though each is mixed with noise at the same SNR. (2). What might account for the variability of results that have been reported in the literature?

The idea that emotionally expressive speech differentially engages a listener's attention has a relatively long history [3]. Based on this idea, Gordon and Hibberts [1] examined the effect of expressing a happy or sad emotion on speech intelligibility in noise (compared with neutral speech as a control). Happy and sad expressions were chosen based on the idea that a happy expression would engender feelings of approach, and so motivate listeners to attend more closely to stimuli, whereas a sad emotion would produce a withdrawal response and so lead listeners to pay less attention. It was found that speech produced with a happy expression was more intelligible than sad speech. Interestingly, it turned out that sad speech was slightly more intelligible than neutral speech, a result that raises a problem for the idea that sad speech would be less intelligible than neutral speech because it exhibits attention discouraging cues.

The results of Gordon and Hibberts [1] are inconsistent with those of a more recent study [4] that examined a larger number of emotions (angry, disgust, fear, happy, sad, surprise and neutral). Dupuis and Pichora-Fuller [4] found that the intelligibility of happy and sad speech did not differ from neutral speech. They did find however that speech expressed with fear was more intelligible than neutral speech. In explaining this pattern of results, Dupuis and Pichora-Fuller also proposed an attention-based account. In this regard, they noted that speech spoken to portray fear was produced with a higher mean F0 than the other emotions. It was suggested that a greater F0 (and/or F0 range) may have contributed to the enhancement of word recognition accuracy by acting as a cue for attentional mechanisms that prioritize speech processing. That is, the F0 cues associated with expressing fear may indicate to a listener that this is an important type of vocal information.

In a very recent study, Gordon & Ancheta [2] used the same happy, sad and neutral stimuli of [1]) and replicated the earlier result: speech spoken to express a Happy affect was more intelligible than speech expressing sadness, which was more intelligible than neutral speech. In a follow-up experiment, the same materials were used except that F0 was manipulated to be constant at 200 Hz. If Dupuis and Pichora-Fuller [4] are correct in proposing that mean F0 is the key acoustic property that leads to differential intelligibility, then adjusting the different speech emotions to have the same (fixed) F0 should eliminate the difference in speech intelligibility in noise. This however was not the case, as the same pattern of results was found even after the F0 manipulation. It would seem then that a relative high F0 (or F0 range) is not the basis for the emotion intelligibility effect.

The proposal that the acoustic cues of certain emotions act to enhance speech processing and so make these utterances more intelligible is an interesting one. However, rather than the intelligibility effect being due to a production style producing cues that attracts a listener's attention, there is a simpler way that acoustic properties could affect intelligibility in noise: by directly affecting audibility. In this regard, it is useful to review the research that has specifically examined how acoustic properties can be influenced by speech style and how such changes relate to intelligibility in noise.

It is well established that the intelligibility of speech in noise differs as a function of speech style. One of the best examples is that Lombard speech (speech produced in noise) is more intelligible than speech spoken in quiet, even when the noise added to both has the same SNR [5; 6; 7]. Compared to speech produced in quiet, Lombard speech typically is longer (especially for sonant sounds), has a higher F0, and has relatively more energy at higher frequencies (flatter spectral tilt), and it has been proposed that the relative increase in intelligibility is due to some combination of these changes [5].

http://dx.doi.org/10.21437/Interspeech.2017-948

The results of a series of experiments by Lu and Cooke [7; 8] make a strong case that a key acoustic change enabling Lombard speech to be more intelligible than quiet speech in the presence of stationary speech-shaped noise is its flatter spectral tilt. For example, Lu and Cooke [7] found that there was a large positive correlation between speech intelligibility and the presence of spectro-temporal glimpses of the target speech (against the noise masker), and that there were more glimpses available for Lombard speech. In a follow-up experiment [8], F0 and spectral tilt were manipulated to investigate which of these had a greater effect on speech intelligibility in noise. It was found that flattening spectral tilt (similar to that found in Lombard speech) played a larger role in improving the intelligibility of a target than did increasing F0 (that did not have a significant effect on intelligibility).

In sum, it would seem that the increase in intelligibility of Lombard compared to quiet speech in the presence of speech-shaped noise is due to a shift in spectral energy to high frequencies in the former. This shift results in more glimpses of the target signal being available to the listener. In addition, this same mechanism appears to account for why clear speech is more intelligible than conversational speech when both are mixed with speech-shaped noise [9].

In our view, then, a straightforward explanation of the differential intelligibility in noise of speech expressing different emotions is similar to that offered for Lombard and clear speech. We examined the plausibility of this proposal by examining the acoustic properties of emotional speech from three different databases (we selected different language databases to capture properties that may be common to emotional expression rather than language). We first determined how the distribution of spectral energy was affected by the expression of emotion, and then calculated a specific glimpse-based metric of the intelligibility of targets in speech shaped noise for each of the emotion conditions. An additional aim of this analysis was to determine the extent to which the emotional speech of different talkers in the databases produced different objective intelligibility scores, since such across talker variation may help to explain why the results of human studies of the intelligibility of different emotions in noise are inconsistent.

2. Method

To determine how different emotional expressions affect the spectral profile of speech we examined three different expressive speech databases. All the databases recorded talkers expressing six emotions plus neutral.

2.1. Databases

The Cantonese Auditory-Visual Expressive speech (CAVES) database [10] contains auditory-visual recordings of 10 native speakers of Cantonese (5 females, mean age = 29.1, SD = 4.9) expressing 50 semantically neutral Cantonese sentences in 7 different emotions (angry, disgust, fear, happiness, sadness and surprise and neutral). The same sentences are produced in each emotion condition. The sentences were selected from the Cantonese Hearing In Noise Test (CHINT) sentences list [11] on the basis that they have a good distribution of tones in each sentence. Sound files were recorded at a sampling rate of 48 kHz.

The Surrey Audio-Visual Expressed Emotion (SAVEE) database [12] consists of auditory-visual recordings of four native English male speakers (postgraduate students and researchers at the University of Surrey, age range 27 to 31 years). The acted emotions express the categories of angry, disgust, fear, happiness, sadness and surprise and neutral. The material consists of 15 TIMIT sentences per emotion: Three sentences in common, two emotion-specific and 10 generic sentences that were different for each emotion and were phonetically-balanced. The three common and 12 emotion-specific sentences were recorded in a neutral style (giving a total 30 neutral sentences) and 120 utterances per speaker. Sound files were recorded at 44.1 kHz.

The Castro and Lima database [13] consists of acted emotions by two women native speakers of European Portuguese. It includes six emotional expressions (angry, disgust, fear, happiness, sadness, and surprise) and neutral speech. The stimulus materials consist of 16 sentences and 16 corresponding pseudo-sentences. Sound files were recorded at 48 kHz sampling rate (16-bit resolution) and normalized for average amplitude.

2.2. Measures of the energy distribution of the spectrum

We used three related measures to characterize speech energy for the different emotion conditions. (1). Center of Gravity (COG, the 1st spectral moment): a measure of the frequency of the entire spectrum expressed in Hz (and weighted by the power spectrum). (2). Spectral Tilt. A measure defined as dB of the first formant (F1) minus dB of the third formant (F3) and expressed as dB/Octave. (3). Spectral slope: Defined as the slope of the long term average spectrum (LTAS) and expressed as the difference in energy (in dB) between the frequency band between 100 and 1000 Hz and the frequency band between 1000 and 4000 Hz (averaging in dB units).

2.3. Objective Intelligibility Metric

We used a glimpse-based objective intelligibility algorithm to calculate an Objective Intelligibility Metric (OIM), [14]. This algorithm is available for download at [15].

In the monaural version of the metric (which is what we used), the target signal and masker are processed by a bank of gammatone filters [16]. The filter centre frequencies boundaries were set at a lower boundary of 100 Hz and upper boundary of 7500 Hz with Equivalent Rectangular Band (ERB) spacing. A 3 dB local SNR threshold was used for defining glimpses (defined as spectro-temporal regions above threshold) and 34 filter bands used. The final metric was calculated by summing glimpse proportion per frequency band weighted by distortion and importance functions; here band importance was set for average speech [see Table 3, 17] and compressed by a quasi-logarithmic function to model ceiling effects. As the aim was to use this metric to examine the plausibility of a bottom-up account of emotion intelligibility differences in noise, it was important to select a noise signal to be used with the algorithm that approximates that used in the human intelligibility studies.

All three studies used 12-talker babble, however the precise makeup the talkers were not given (e.g., talker's sex and age). For simplicity, we used unmodulated speech-shaped noise (SSN) based on the LTAS of each talker. It should be noted that Simpson and Cooke [18] have shown that for at least for consonant identification, 12-talker babble is likely to be a more effective masker than unmodulated speech-shaped noise. Given this we used a range of SNRs for target and noise (between -2 and -6 SNR) when calculating the OIM.

3. Results

We first determined how emotional expression affects the distribution of spectral energy by determining some standard measures for the three databases. We measured COG, Spectral Tilt and Spectral Slope and compared these measures from the emotion conditions with those of the neutral speech condition. We then determined the OIMs for the same emotion conditions when mixed with SSN at 5 different SNR (-6, -5, -4, -3, -2).

Figure 1 shows both the spectral energy measures and the OIM (averaged over the SNR levels) for the Cantonese database (averaged over talkers). Overall, the conditions Angry, Happy and Surprise had higher COG scores and flatter spectral tilt and slope compared to neutral condition. Also, there is a clear correspondence between the COG scores and the OIM ones. Indeed, of all the energy measures, COG had the largest Pearson correlation with OIM scores, r = 0.5 (averaged over the separate SNR levels).

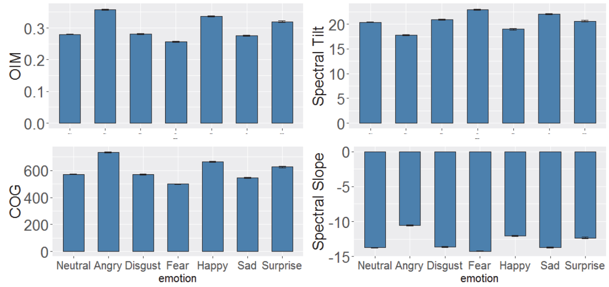

Figure 1: *OIM scores and the three spectrum energy measures for the Cantonese Emotion database.*

The effect of the emotion conditions on intelligibility (measured against the neutral condition) was determined by using a generalized linear mixed model applied to the OIM scores. We used emotion condition as the fixed effect, and the intercepts for speaker and item as random effects. P-values were obtained by conducting simultaneous tests that were specified by a matrix of contrasts across the conditions using the glht function of the lme4 R package [19]. The outcomes of the contrast of each emotion condition with the neutral condition are shown in Table 1.

Table 1: *Generalized linear mixed model contrasts for the Cantonese database*

Linear Hypotheses:	Estimate	SE	z value	Pr(>|z|)
Angry vs. neutral	0.0819	0.0006	130.67	<0.001
Disgust vs. neutral	0.0052	0.0006	8.275	<0.001
Fear vs. neutral	-0.0190	0.0006	-30.303	<0.001
Happy vs. neutral	0.0606	0.0006	96.673	<0.001
Sad vs. neutral	0.0017	0.0006	2.691	0.0358
Surprise vs. neutral	0.0448	0.0006	71.382	<0.001

As can be seen in Table 1, according to the OIM scores, the Angry, Disgust, Happy and Surprise conditions were more intelligible than neutral one; the Sad condition did not differ from neutral and the Fear condition was less intelligible than neutral.

Following on from the finding that holding F0 constant did not affect human intelligibility scores [2], we adjusted the F0 for all the audio files in the Cantonese database (set to a constant 130 Hz for the male talkers and 224 Hz for the female talkers) and then re-calculated OIM scores. As it turned out the pattern of results was very similar, except that "Disgust vs. neutral" contrast was no longer significant (Z = 1.320, p = 0.597) and "Sad vs. neutral" now was (Z = 4.271, p < 0.001).

Figure 2 shows both the spectral energy measures and the OIM (averaged over the SNR levels) for the SAVEE database (averaged over talkers).

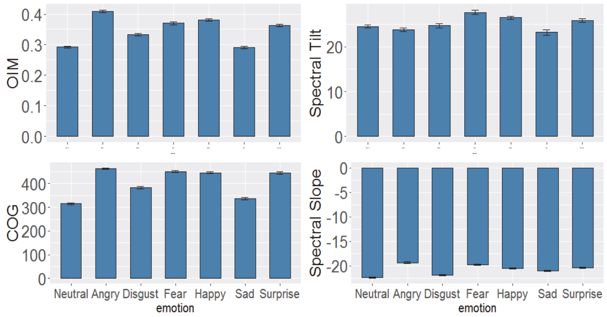

Figure 2: *OIM scores and the three spectrum energy measures for the SAVEE Emotion database.*

The COG results showed that Angry, Disgust, Fear, Happy and Surprise (and possibly Sad) had scores higher than neutral. Once again, the OIM scores showed the same pattern as the COG scores, and the COG scores had the highest correlation with the OIM scores (r = .53).

The effect of the emotion conditions on intelligibility (versus the neutral condition) was determined by a generalized linear mixed model applied to the OIM scores. Emotion condition was the fixed effect and the intercepts for speaker and item were treated as random effects. The outcomes are shown in Table 2.

Table 2: *Generalized linear mixed model contrasts for the SAVEE database*

Linear Hypotheses:	Estimate	SE	z value	Pr(>|z|)
Angry vs. neutral	0.1166	0.0113	10.321	<0.001
Disgust vs. neutral	0.0416	0.0113	3.556	<0.001
Fear vs. neutral	0.0785	0.0113	6.951	<0.001
Happy vs. neutral	0.0897	0.0113	7.939	<0.001
Sad vs. neutral	-0.0010	0.0113	-0.088	1.0000
Surprise vs. neutral	0.0710	0.0113	6.279	<0.001

Based on Table 2, it would appear that all the emotion conditions except for Sad were more intelligible than neutral. In this regard, the biggest difference with the results based on the Cantonese database is that the Fear condition has swapped from being less to more intelligible than neutral.

The spectral energy measures and OIM scores for the Castro and Lima database are shown in Figure 3.

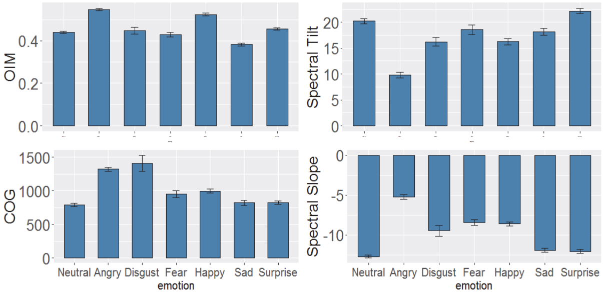

Figure 3: *OIM scores and the three spectrum energy measures for the* Castro and Lima *Emotion database.*

For this database, the patterning of COG and OIM scores is not as close as the other two databases. Indeed, the correlation between the COG and OIM scores was modest (r = 0.26) and the highest correlation with OIM was with the Spectral Slope scores (r = 0.55). It is nevertheless clear that the OIM scores did differ across the different emotion conditions. We examined the effect of the emotion conditions on intelligibility (measured against the neutral condition) by once again modelling the OIM scores with a generalized linear mixed model (see Table 3).

Table 3: *Generalized linear mixed model contrasts for the Castro and Lima database*

Linear Hypotheses:	Estimate	SE	z value	Pr(>|z|)
Angry vs. neutral	0.1166	0.0190	5.770	<0.001
Disgust vs. neutral	0.0416	0.0255	0.409	0.998
Fear vs. neutral	0.0785	0.0205	-0.370	0.999
Happy vs. neutral	0.0897	0.0195	4.394	<0.001
Sad vs. neutral	-0.0010	0.0191	-2.825	0.025
Surprise vs. neutral	0.0710	0.0190	0.942	0.867

As can be seen from Table 3, only two emotion conditions were more intelligible than neutral (Angry and Happy) and one emotion was less intelligible than neutral (i.e., Sad). It is clear from a consideration of all three databases that although stimuli in the Angry and Happy conditions were always more intelligible than those of the neutral condition, the pattern for the other emotions was variable.

Not only was there variation in pattern of the OIM scores across the three databases, but there was also variation across the talkers within a database. Figure 4 shows the OIM scores for all of the ten talkers of the Cantonese database.

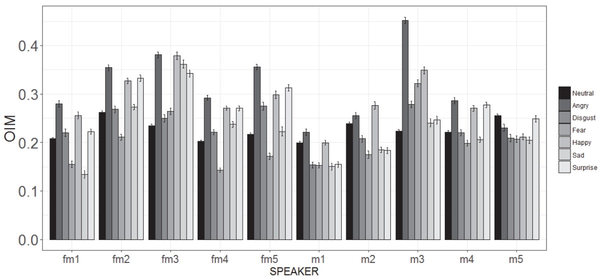

Figure 4: Variation in *OIM scores across talkers in the Cantonese database for target/noise SNR -6.*

We examined the pattern of OIM scores across the emotion conditions as a function of emotion rater agreement (high versus low) to see if some of the variation in the OIM scores could be explained by how well the emotion was expressed (these data were available only for the Cantonese and Castro and Lima datasets). For the Cantonese database, the largest difference in OIM scores for the high versus low agreement stimuli was for Angry and Happy, and for the Castro and Lima database, Disgust, Happy and Surprise. Interestingly, these emotions were the ones that showed a clear OIM advantage against neutral (suggesting a consistency in their pattern), so it is not obvious that the quality of the emotion expression can explain the variation in OIM scores.

4. Discussion

The results of the spectral analyses were consistent with the well-established idea that one property in the signalling of different emotions involves the distribution of energy in the speech frequency spectrum [20]. In general, across all three emotion databases, we found that speech expressing the emotions Happy and Angry had relatively more energy in higher frequencies compared to neutral speech. Similar shifts in the spectrum of speech expressing emotions like Disgust, Fear and Surprise were found but occurred only in some of the databases.

Shifts in the balance of speech spectral energy, especially that which involves a relocation of energy from low to higher frequency regions, will have an impact on the intelligibility of speech when mixed with speech-shaped noise due to more speech energy becoming available to the listener. This was modelled here with an OIM and the results showed Happy and Angry to be the most intelligible. Within and across the databases there was variation in which other emotions were found to be more intelligible than neutral speech. This variability likely reflects a lack of clear constraint in how emotional speech is produced and realized (e.g., Fear could be understated or dramatic) and could help explain the different results that obtain in the human intelligibility studies.

Finally, it is worth pointing out that cognitive factors (e.g., attentional effects) may also play a role in why speech expressed with some emotions is more intelligible in noise. However, we suggest that any examination of this interesting idea first needs to carefully consider how the production of vocal emotions affects speech audibility and to build this consideration into the design of a study.

5. Acknowledgements

The authors acknowledge support from two ARC Discovery grants, DP130104447 and DP50104600.

6. References

[1] Gordon, M. S., & Hibberts, M. (2011). Audiovisual speech from emotionally expressive and lateralized faces. The Quarterly Journal of Experimental Psychology, 64(4), 730-750.

[2] Gordon, M. S., & Ancheta, J. (2017). Visual and acoustic information supporting a happily expressed speech-in-noise advantage. The Quarterly Journal of Experimental Psychology, 70(1), 163-178.

[3] Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: brain responses to angry prosody in meaningless speech. Nature neuroscience, 8(2), 145-146.

[4] Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of emotional speech in younger and older adults. Ear and hearing, 35(6), 695-707.

[5] Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. The Journal of the Acoustical Society of America, 93(1), 510-524.

[6] Pittman, A. L., & Wiley, T. L. (2001). Recognition of speech produced in noise. Journal of Speech Language and Hearing Research, 44(3), 487-496.

[7] Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. The Journal of the Acoustical Society of America, 124(5), 3261-3275.

[8] Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Communication, 51(12), 1253-1262.

[9] Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. The Journal of the Acoustical Society of America, 115(1), 362-378.

[10] Chong, C.S., Kim, J. & Davis, C. (2014). Development of an Audiovisual Cantonese Emotional Speech Database, 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (oCOCOSDA).

[11] Wong, L. L., & Soli, S. D. (2005). Development of the Cantonese hearing in noise test (CHINT). Ear and hearing, 26(3), 276-289.

[12] SAVEE database http://kahlan.eps.surrey.ac.uk/savee/

[13] Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. Behavior Research Methods, 42(1), 74-81.

[14] Tang, Y., Cooke, M., Fazenda, B. M., and Cox, T. J. (2016). "A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers," Journal of the Acoustical Society of America, 140(3), 1858-1870.

[15] https://dx.doi.org/10.17866/rd.salford.3549774

[16] Patterson, 1988 Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (1988).“SVOS Final Report: The Auditory Filterbank,” Technical Report 2341, Medical Research Council (MRC) Applied Psychology Unit.).

[17] ANSI S3.5-1997 ANSI (1997). S3.5, “Methods for the calculation of the Speech Intelligibility Index” (Acoustical Society of America, New York).

[18] ref Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. The Journal of the Acoustical Society of America, 118(5), 2775-2778.

[19] Bates D., Mächler M., Bolker B., Walker S. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Software. 67, 1–48.

[20] Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression, J. Pers. Soc. Psychol, 70(3), 614-636.