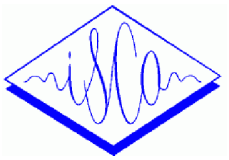
Explaining the visual and masked-visual advantage in speech perception in noise: The role of visual phonetic cues

Vincent Aubanel, Chris Davis, Jeesun Kim

ISCA Archive
http://www.isca-speech.org/archive



FAAVSP - The 1st Joint Conference on
Facial Analysis, Animation, and
Auditory-Visual Speech Processing
Vienna, Austria,
September 11-13, 2015

Explaining the visual and masked-visual advantage in speech perception in noise: The role of visual phonetic cues

Vincent Aubanel, Chris Davis, Jeesun Kim

The MARCS Institute, University of Western Sydney, Australia
v.aubanel@uws.edu.au

Abstract

Visual enhancement of speech intelligibility, although clearly established, still resists a clear description. We attempt to contribute to solving that problem by proposing a simple account based on phonetically motivated visual cues. This work extends a previous study quantifying the visual advantage in sentence intelligibility across three conditions with varying degrees of visual information available: auditory only, auditory visual orally masked and auditory-visual. We explore the role of lexical as well as visual factors, the latter derived from groupings in visemes. While lexical factors play an undiscriminative role across modality conditions, some measure of viseme confusability seems to capture part of the performance results. A simple characterisation of the phonetic content of sentences in terms of visual information occurring exclusively inside the mask region was found to be the strongest predictor for the auditory-visual masked condition only, demonstrating a direct link between localised visual information and auditory-visual speech processing performance.

Index Terms: Auditory-visual speech processing, visemes, sentence intelligibility, visual advantage

1. Introduction

It has long been established that seeing the talker's face contributes dramatically to speech intelligibility in noise [1], which can be quantified as an equivalent intensity increase of about 11 dB [2]. As highlighted in a recent review of auditory-visual speech processing however [3], little is known about the mechanisms at play in that advantage. In what follows we briefly review some approaches to index visual speech information.

One indirect method for probing the nature and extent of visual speech information has been used in a series of studies quantifying the extent to which speech acoustics can be predicted by associated visual facial movements. This is an indirect method, as it assumes that listeners make use of the information available from visible speech production gestures as well as from the patterning of these with other non-visible articulatory gestures. For example, multivariate correlation analyses found that 7 two-dimensional vocal tract markers (including the jaw and the lips) accounted for 80% of the face motion behaviour as measured with 18 three-dimensional markers on the face [4, 5]. Those studies also showed a similarly good prediction of the speech acoustics from the vocal tract configuration, especially in the F_2 region which is closely related to the volume of the frontal oral cavity. Rigid head motion was found to be strongly correlated with f_0, a functional rather than structural link as evidenced by talker's ability to consciously decorrelate the two. Rigid head motion information enhances speech intelligibility [6] with the tentative explanation for this being that it does so by helping the perceiver focus on suprasegmental rather than the segmental level.

A more direct approach to assessing visual speech information comes from occlusion studies, e.g., [7]. Such studies have explored the influence of different facial regions and highlighted the importance of oral regions in enhancing visual speech intelligibility. Conversely, studies that have measured gaze patterns to identify information bearing regions suggest that the talkers' eye regions are more attended to than the mouth regions [8], even when increasing the talker's face projection to render simultaneous viewing of the eye and mouth regions impossible within the fovea [9].

Limiting the visual information to the lips provides some cues to retreive associated speech acoustics when presented in noise, but the intelligibility benefit is removed when only timing information is retained, by converting the lip rounding gesture to an equivalent rectangle height [10]. A complementary approach was adopted by [11] who investigated the relative contribution of timing and content information in speech, by superimposing a circular patch on the oral region: the full-face is proposed to contain both timing and form information while only timing information is present in the masked face. They found that timing information alone was beneficial in discriminating speech from non-speech over a baseline of a static face. An additional facilitatory effect was observed when form information was present only for speech stimuli, suggesting a cross-modal priming associating the matching spectral characteristics of the speech stimuli and the form of lips and mouth.

The visual equivalent of phonemes, visemes [12], have for a long time been a focus of research, with the hope that a categorisation of visual features of speech sounds would provide some explanatory power to describe and understand auditory-visual speech perception [13, 14, 15, 16]. While linguistic-based rather than data-driven groupings seem to perform best, to date this approach has not been successful in identifying a set of phoneme-to-viseme mapping which would exceed what can be achieved with a phonemic categorisation, even when accounting for contextual articulation of phonemes in establishing mappings (e.g., [17]). Often the main conclusion is that phonemes, not visemes, should remain the basic unit for *visual* speech recognition [16].

In this paper we explore an approach for explaining the visual advantage in auditory-visual speech perception that follows on from the masking procedure used by [11]. This work uses data obtained in a previous study [18] that showed a clear increasing performance with increasing visual area available in speech perception in noise. Glimpsing of auditory information [19] was found to be a strong predictor for performance, with the strongest correlation occurring in the auditory only condition, suggesting that in the absence of visual information, par-

ticipants relied more on bottom-up cues. A signal-based evaluation of visual movement was not a strong predictor of performance in visual conditions, leaving the visual advantage lacking a clear explanation. We attempt to fill this gap here, by focussing on top-down cues such as lexical factors, the visibility of various viseme groupings, and a phonetically motivated quantification of visual information restricted to the oral region.

The remainder of the paper is divided as follows: in Section 2 we present the corpus on which the study builds, as well as a variety of lexical and visual predictors. Section 3 shows results of a correlation analysis with sentence intelligibility which are discussed in Section 4.

2. Methods

2.1. Data

The data analysed here is taken from study reported in [18]. In that study, participants were presented with audio-visual productions of IEEE sentences [20] by a native Australian female speaker in three auditory-visual conditions, varying in amount of visual information made available. In all conditions, the auditory channel was mixed with speech-shaped noise whose spectrum approximated the long-term averge spectrum of the talker, at a signal-to-noise ratio of -3 dB. The three visual conditions were: Auditory only (hereafter, **A**), where a grayscale static image of the talker face was presented with closed lips. The image was presented in grayscale to make it clear to the participants that they should not expect any visual information. In the Auditory-Visual masked condition (**AVm**), the video recording of the talker was played synchronously with the auditory channel. The region of the video was limited to the lower part of the face and a circular patch with a fixed radius determined to mask the lip area was superimposed on the video. The color of the patch was gray with a saturation value that matches the average value of the masking region across the corpus. The Auditory-Visual condition (**AV**) was identical to the AVm condition with the exception of the superimposition of the circular patch. Figure 1 summarizes the conditions.

The participants' task was to type the words that they heard and their response was scored automatically by counting the number of correct keywords recognised, accounting for homophones and common spelling mistakes.In the following, we take the dependent variable as the proportion of keywords correctly recognised.

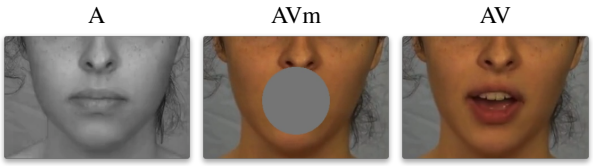


Figure 1: *The three auditory-visual conditions used in [18]. Color indicate animated video.*

2.2. Lexical and visual information

Following the analyses of [18], we set out to explain the visual over visual masked advantage with a range of lexical predictors, with the idea that lexical knowledge could be recruited differently for the visual modality. Lexical predictors were obtained for each word of the sentences by querying the En-

Table 1: *Grouping of phonemes into visemes and their associated visibility ranking, when available. The rightmost column shows the coding of exclusively orally salient feature (XOS), with grayed rows for phonemes falling in that category.*

Phon.	IPA	Group (visib.) Jeffers [21]	Group Hazen [22]	Visib. Ramage [16]	XOS
Stop					
p	p	C (3)	LB	7	no
b	b	C (3)	LB	5	no
t	t	J (10)	SB	4	no
d	d	J (10)	SB	4	no
k	k	K (11)	SB	2	no
g	g	K (11)	SB	8	no
Affricate					
ch	ʧ	F (6)	Pal	12	yes
jh	ʤ	F (6)	Pal	9	yes
Fricative					
f	f	A (1)	LFr	7	yes
v	v	A (1)	LFr	6	yes
th	θ	E (5)	SB	13	yes
dh	ð	E (5)	SB	1	yes
s	s	H (8)	AlCl	1	no
z	z	H (8)	AlCl	1	no
sh	ʃ	F (6)	Pal	9	yes
zh	ʒ	F (6)	Pal	8	yes
hh	h	K (11)	FV	1	no
Nasal					
m	m	C (3)	LCl	4	no
n	n	J (10)	AlCl	3	no
ng	ŋ	K (11)	VlCl	14	no
Lateral/Glide					
w	w	B (2)	RV	1	no
r	ɹ	B (2)	R	2	no
l	l	J (10)	L	2	yes
y	j	I (9)	Y	5	yes
Diphthong					
ey	eɪ	I (9)	FV	5	no
oy	ɔɪ	G (7)	RV	10	no
aw	aʊ	D (4)	RV	6	no
ay	aɪ	I (9)	FV	2	no
Vowel					
iy	iː	I (9)	OV	3	no
uw	uː	B (2)	RV	10	no
ih	ɪ	I (9)	OV	2	no
uh	ɝ	B (2)	RV	11	no
ow	oʊ	B (2)	RV	1	no
ax	ə	I (9)	OV	1	no
ah	ʌ	I (9)	BV	3	no
eh	ɛ	I (9)	FV	1	no
er	ɜ	B (2)	R	3	no
ae	æ	I (9)	FV	2	no
aa	ɑ	I (9)	BV	3	no
ao	ɔː	I (9)	RV	4	no

glish Lexicon Project database[1] [23]. This consisted of obtaining the length of the word in phonemes (NPhon) and syllables (NSyll), the log frequency in a corpus of written American English (Log_Freq_HAL) and the phonological neighborhood size (Phono_N_H). Lexical predictors were computed for each sentence as the mean value of the predictor of the keywords of that sentence, i.e., the function keywords were not scored.

Next, we explored different groupings of phonemes according to their visual saliency, or visibility. We used two commonly used phoneme-to-viseme maps, the linguistically motivated map of Jeffers & Barley [21] (hereafter, Jeffers) where phonemes are grouped into the same viseme class when their articulation should make them appear similar (for example, /p/ and /b/ are grouped together as they both involve a lip protrusion gesture). This mapping contains 11 categories (plus a silence category not used here), each assorted with a visibility rank. This coding, as well as other following codings, are reported in Table 1. Another widely used mapping is that proposed by Hazen et al. [22], established in a data-driven approach by clustering phonetically-annotated visual frames. The confusion intra- and inter-category was evaluated by Ramage [16, Table 6.4]. Based on this work, Ramage also proposed a visibility ranking of phonemes. Finally, in order to test the visual over the masked-visual advantage, we also made a simple grouping into one of two categories, according to whether the phoneme's articulation would provide visual information mainly in the oral region. We excluded from this category two different sound types. The first constists of speech sounds thought to have low visibility such as the lax vowel /ɪ/, which are unlikely to constitute a target for articulators but rather their visual appearance is expected to be influenced by neighboring more salient (and/or accented) phonemes. Second, we also excluded from this category those phonemes whose articulation is marked enough so that their visible manifestation affects a broader region than just the lip area, and may be predictable from movement of peri-oral articulation, such as the jaw motion in the case of articulation of the stressed /ɑ/ or the marked lip protrusion in the articulation of /w/. We call this indicator the *exclusively orally salient* feature (XOS).

For continous predictors, we computed a value per sentence as the mean score of the parameter across keywords of that sentence. For XOS, the value is taken as the proportion of phonemes of that class across all phonemes of keywords of the sentence.

3. Results

Table 2 shows the correlation coefficients for each predictor with mean sentence score across subjects in each modality condition.

3.1. Lexical predictors

Starting with lexical predictors, one can observe that the length of keywords has no effect on sentence intelligibility when measured in number of phonemes [all $p > .1$]. The number of syllables of keywords was however found to be associated with sentence intelligibility for the AV condition only [$p = .02$], with longer words being better recognised than shorter words. It should be noted however that on this sentence material with a constrained structure, the mean number of syllable can take only one of three values as a result of the total number of syllables across all sentences being either 5, 6 or 7, with 167 out of 180 sentences (92.8%) having exactly 6 syllables. Therefore, a correlation analysis is likely to be influenced by the score of the remaining small number of sentences falling outside of this group.

[1] http://elexicon.wustl.edu. Last checked 1st May 2015

Table 2: *Correlation coefficient of lexical and visual predictors with mean sentence score for each visual modality condition.* p-*values given by Pearson's product-moment correlation test are indicated according to classical coding of significativity levels.*

Predictor	A		AVm		AV	
Lexical						
NPhon	−0.062		−0.019		−0.066	
NSyll	0.111		0.130	.	0.173	*
Log_Freq_HAL	0.285	***	0.293	***	0.310	***
Phono_N_H	−0.038		−0.060		−0.065	
Visual						
Jeffers vis.	−0.045		−0.036		−0.052	
Hazen NInter	−0.098		−0.190	*	−0.121	
Hazen NIntra	0.050		−0.039		−0.057	
Ramage vis.	0.165	*	0.163	*	0.110	
XOS prop.	−0.143	.	−0.226	**	−0.073	

The keywords' log frequency was found to be robustly correlated with sentence intelligibility [all $p < .001$], however there was no difference across modality conditions. Phonological neighborhood did not yield any significant correlation with performance [all $p > .1$] across the three conditions. Taken together, lexical factors seem to play little or no role in explaining the visual advantage in perceiving speech in noise.

3.2. Visual predictors

Average keyword visibility as measured by Jeffers phoneme-to-viseme mapping and its associated visibility rank did not correlate with sentence intelligibility [all $p > .1$]. While intra-viseme confusion with the Hazen mapping did not have explanatory power either, inter-viseme confusion was negatively correlated with performance for the AVm condition only [$p = .01$], indicating that sentences with a higher proportion of phonemes easily confused with another's viseme group seem to carry relevant information related to orally masked speech.

Another notable result concerned the Ramage visibility ranking, for which sentences containing a higher proportion of visually 'clean' phonemes received higher scores for the A [$p = .027$] and AVm [$p = .029$] conditions. The proportion of such phonemes do not seem to provide an extra benefit when the full lower face was visible.

The strongest predictor of masked visual performance is obtained by the XOS proportion, and is also represented in Figure 2. Sentences with a high proportion of phonemes which present information mainly in the oral region were the main predictor of masked visual intelligibility [AVm: $p = 0.003$].

4. Discussion

We have shown that while lexical factors appear to play little role in explaining the visual advantage in perceiving sentences in noise, a simple characterization of visual phonetic information is able to predict the amount of information listeners attend to in retreiving speech information from orofacial movements.

The fact that high frequency words are more easily recognised in noise is long established [24]. As was also found by

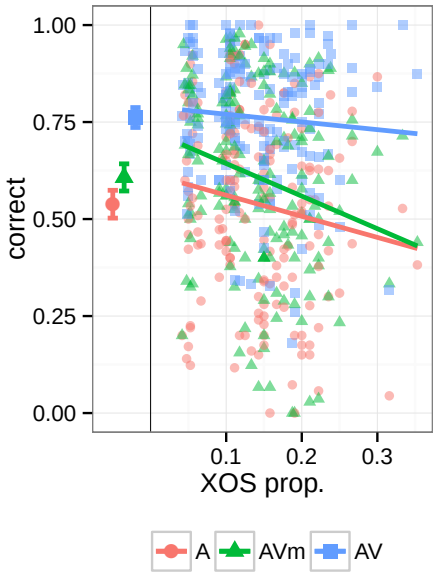


Figure 2: *Mean sentence performance as a function of XOS phoneme proportion for the three modality conditions, with regression line overlaid. Mean performance across conditions found in the [18] study are reported on the left of the Figure, with error bars showing 95% confidence intervals ($N = 27$).*

[17], the current study shows that this effect is not modulated by the availability of visual information. Although it is difficult to speculate further given the distributional constraints of the sentence material employed here, the result that additional context, as measured by number of syllables, has a greater role in the full- face condition is an interesting one, pointing to a nonlinear combination of auditory and visual cues in speech intelligibility. Systematic variation of number of syllables in auditory only condition shows an effect on intelligilibity, e.g., [25], therefore it is expected that the observed effect would be even higher in an audiovisual condition (see also [26]).

Visemic grouping of phonemes and their associated visibility scores did not make a clear contribution in explaining performance difference between orally masked and full-face visual speech. When there was an effect, such as the positive correlation for the Ramage visibility predictor, reasons may be found in the phonetic characteristics of the sounds: visually 'clear' phonemes (such as /ŋ/) may also lead to similar 'clear' phonetic realisation. Similarly, increasing inter-category confusability using the Hazen grouping may be associated with phonetic confusability.

Following that observation, it is interesting to note that 8 out of 10 of the phonemes selected as orally salient are fricatives or affricates, for which the spectral and amplitude envelope characteristics are, relatove to other phonemes, likely to be masked by the type of noise employed here. This may in turn explain the hint of negative correlation with intelligibility in the auditory only modality.

In all, the results reported here point to visual cues that are easily described in articulatory terms. This description can account for the visual advantage difference between full-face and masked face perception of audiovisual speech.

5. Acknowledgements

The research leading to these results was partly funded by the Autralian Research Council under grant agreement DP130104447.

6. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[2] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise." *Brit. J. Audiol.*, vol. 21, no. 2, pp. 131–141, May 1987.

[3] E. Vatikiotis-Bateson and K. G. Munhall, "Auditory-visual speech processing: Something doesn't add up," in *The Handbook of Speech Production*, M. A. Redford, Ed. John Wiley & Sons, 2015, pp. 178–199.

[4] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *J. Phon.*, vol. 30, no. 3, pp. 555–568, Jul. 2002.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1–2, pp. 23–43, 1998.

[6] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility head movement improves auditory speech perception," *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[7] S. M. Thomas and T. R. Jordan, "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception." *J. Exp. Psychol. Human.*, vol. 30, no. 5, pp. 873–888, Oct. 2004.

[8] C. R. Lansing and G. W. McConkie, "Attention to facial regions in segmental and prosodic visual speech perception tasks," *J. Speech Lang. Hear. R.*, vol. 42, no. 3, pp. 526–539, 1999.

[9] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisual speech perception." *Percept. Psychophys.*, vol. 60, no. 6, pp. 926–940, Aug. 1998.

[10] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, Sep. 2004.

[11] J. Kim and C. Davis, "How visual timing and form information affect speech and non-speech processing." *Brain Lang.*, vol. 137, pp. 86–90, Oct. 2014.

[12] C. G. Fisher, "Confusions among visually perceived consonants," *J. Speech Lang. Hear. R.*, vol. 11, no. 4, pp. 796–804, 1968.

[13] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-Visual Speech Recognition," Center for Language and Speech Processingm, The John Hopkins University, Baltimore, MD, USA, Tech. Rep., 2000.

[14] S. Hilder, B.-J. Theobald, and R. Harvey, "In pursuit of visemes," in *AVSP*, Jul. 2010, pp. 8–2.

[15] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for visual speech recognition," in *ICPRAM*, 2012, pp. 322–329.

[16] M. D. Ramage, "Disproving Visemes As The Basic Visual Unit Of Speech," Ph.D. dissertation, Curtin University, Dec. 2013.

[17] S. Mattys, L. E. Bernstein, and E. T. J. Auer, "Stimulus-based lexical distinctiveness as a general word-recognition mechanism." *Percept. Psychophys.*, vol. 64, no. 4, pp. 667–679, May 2002.

[18] J. Kim, V. Aubanel, and C. Davis, "The effect of auditory and visual signal availability on speech perception," in *ICPhS*, Glasgow, UK, 2015.

[19] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[20] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, M. Weistock, V. E. McGee, U. P. Pachl, and W. D. Voiers, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. Audio Acoust.*, pp. 225–246, 1969.

[21] J. Jeffers and M. Barley, *Speechreading (Lipreading)*. Springfield, IL, USA: Charles C Thomas, 1971.

[22] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer," in *ICMI '04*. New York, New York, USA: ACM Press, 2004, pp. 235–242.

[23] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English Lexicon Project," *Behav. Res. Meth.*, vol. 39, no. 3, pp. 445–459, Aug. 2007.

[24] D. Howes, "On the Relation between the Intelligibility and Frequency of Occurrence of English Words," *J. Acoust. Soc. Am.*, vol. 29, no. 2, pp. 296–305, 1957.

[25] H. Rubenstein, L. Decker, and I. Pollack, "Word Length and Intelligibility," *Lang. Speech*, vol. 2, p. 175, 1959.

[26] M. Dubois, D. Poeppel, and D. G. Pelli, "Seeing and Hearing a Word: Combining Eye and Ear Is More Efficient than Combining the Parts of a Word," *PLoS ONE*, vol. 8, no. 5, p. e64803, May 2013.