SPEAKING UNDER COVER: THE EFFECT OF FACE-CONCEALING GARMENTS ON SPECTRAL PROPERTIES OF FRICATIVES

Natalie Fecher & Dominic Watt

Department of Language and Linguistic Science, University of York, York, UK
natalie.fecher@york.ac.uk; dominic.watt@york.ac.uk

ABSTRACT

This paper firstly reports on the design of an audio-visual 'face cover' corpus. High-quality audio and video recordings were taken of 10 speakers reading phonetically-controlled stimuli under various face disguise conditions. Possible articulatory, acoustic and perceptual effects of the masks in a forensic context are introduced. Secondly, preliminary results of a spectral analysis of voiceless fricatives, taken from a subset of speakers, are presented.

Keywords: forensic phonetics, acoustics, disguise, fricatives, spectral moments

1. INTRODUCTION

Forensic speech science (FSS) depends, among other things, on information about a speaker's speech and language gathered from questioned audio or video material, the quality and quantity of which is often limited. To account for discrepancies between real-life casework and the empirical research that underpins it, FSS experts are increasing their efforts to simulate more realistic scenarios when conducting experiments. A three-year project has been initiated as part of the interdisciplinary Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics' [1], focusing on the influence of forensically-relevant face-concealing garments[1] (henceforth FCGs) on the physiological, acoustic and linguistic levels of the speech chain. The project investigates both human and machine performance during speech and speaker recognition tasks under visually and acoustically degraded conditions.

Where an FCG obstructs the talker's face we could plausibly anticipate effects in three domains:

1.1. Misarticulation and compensation

In the first domain (speech production), we might expect FCGs to interfere in various ways with speech articulation. Firstly, misarticulations could be attributed to physiological and somatosensory effects, such as lip/nose contact, restricted jaw elevation [5] and skin stretching [4, 10]. As the same FCGs were used for all speakers in this study, the subjects' head sizes determined how tightly some FCGs were attached to the speakers' faces and articulators. Simultaneously, each subject might reveal idiosyncratic articulatory compensation strategies, like an increase in vocal effort [2]. For certain FCGs, compensatory phenomena could also result from the speakers' ears being covered, impairing auditory self-monitoring. As physio-logical and acoustic events in the vocal tract are interdependent, effects in this domain will alter the acoustic signal. For instance, a perturbed lip protrusion in [ʃ] due to a mask being in contact with the speaker's lips may shorten the front tube of the vocal tract, leading to frequency shifts.

1.2. Acoustic damping effects

The different mask materials will modify the acoustic properties of the signal by affecting the sound transmission/absorption characteristics to varying degrees. The FCGs are assumed to act like a low-pass filter, attenuating the level of sound energy in higher frequency bands [2, 8, 17].

1.3. Impaired recognition and visual speech

FCGs impose a new level of complexity in the listener's search for perceptual cues in the signal. We propose that the factors in (i) and (ii) impose significant cognitive demands in audio-visual (AV) speech perception. Impaired intelligibility of AV stimuli is anticipated because of interference with speech production and the acoustic signal [8, 17], compounded by impoverished visual cues (see e.g. [12, 13, 14]). In this paper we report on the design of an AV database as well as preliminary findings from the acoustic domain, i.e. the influence of FCGs on the spectral properties of voiceless fricatives.

2. METHOD

2.1. The AV 'face cover' corpus

10 speakers (5 M, 5 F, age 21-36) were recorded in a professional TV studio at the University of York. No participant had a history of impaired speech, hearing or vision, and none had experience of regularly wearing FCGs (e.g., for recreational, occupational or religious reasons). All were native British English speakers with training in phonetics, enabling them to reliably produce the target stimuli presented using IPA symbols. Their task was to read aloud a list of 64 nonsense /C_1ɑC_2/ syllables embedded in the carrier phrase *He said* <*stimulus*>. 18 English consonants, i.e. /p t k b d g f s ʃ θ v z ʒ ð m n ŋ h/, occurred twice in each syllable position (onset/coda). Wearing various FCGs, the speakers read the list a total of nine times (see Table 1). The order of stimuli, stimuli lists, and guise conditions was randomised for each subject to mitigate fatigue effects. High-quality audio recordings were captured using three microphones: one headband and two shotgun microphones placed in front of and behind the speaker. Footage of the subjects' head and shoulders was filmed from two camera angles (face-on/half-profile). In total, 6,120 consonant utterances were recorded per micro-phone (10 speakers * 9 coverings[2] * 18 consonants * 2 repetitions[3] * 2 syllable positions).

2.2. Face-concealing garments selected

The selection criteria for the face coverings were *forensic relevance*, *mask material* and *parts of the face concealed*.

Table 1: Control condition and 7 FCGs, incl. the relevant material that covers the speaker's mouth/nose.

Some guises are used for the commission of crimes like robberies, assaults or terrorist activities (HEL/HOO/BAL/RUB). In most cases they serve to change a person's *visual* appearance, rather than to deliberately disguise the *voice* for the purpose of concealing identity [18]. Others are worn for religious (NIQ) or safety/security purposes (HEL/SUR). All could possibly lead to miscommunication or complaints based on degraded intelligibility [8].

2.3. Speech material

The choice to investigate fricatives for the acoustic analysis presented in this paper was motivated by their perceptual confusability [9, 11], their relevance as consonantal features in FSS, and an anticipated larger attenuation by certain FCGs of energy in higher frequency bands that are particularly discriminative for this phoneme class [8]. The spectral properties of fricatives are defined by the place, degree and shape of the narrowest constriction in the vocal tract, especially the length of the front tube, and marginally by pressure and rate of airflow [6]. Of particular interest for FSS is the intra- and inter-speaker variability found when specifying spectral peak and shape information. Analysis of individual utterances shows substantial inter-speaker overlap such that speaker A's [s] is acoustically indistinguishable from speaker B's [ʃ] production [5].

The speech material included in the present analysis consists of two tokens per syllable position of the voiceless sibilants /s ʃ/ and non-sibilants /f θ/ taken from the recordings of 2 female and 2 male speakers. Segmentation points were determined based on visual inspection following established procedures (see e.g. [5, 7]).

2.4. Acoustic measurements

Digital recordings (48kHz, 16-bit) were made using an omnidirectional headband microphone (DPA 4066) placed ca. 3cm from the right corner of the speaker's mouth. Measurements were taken automatically from wideband spectrograms (Gaussian, window length 5ms) in *Praat 5.1.44*. Following [16], no pre-emphasis filter (other than the *Praat* default of 6dB/oct) was implemented. Averaged FFT power spectra were computed over non-filtered speech, thus taking the frequency range up to 24kHz into account, which is beneficial given that fricatives can carry place in-formation above the classic 10kHz cutoff [15, 16].

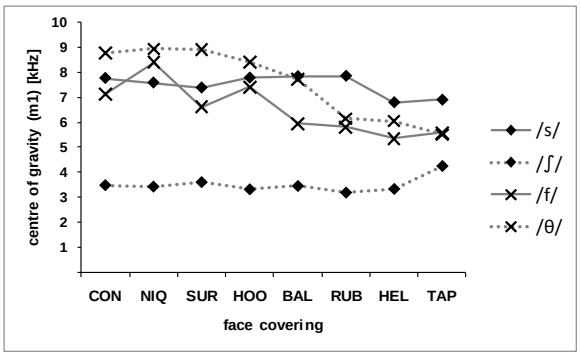
Despite increasing numbers of acoustic studies of obstruents, the set of quantitative parameters to characterise the acoustic structure of fricatives still lacks standardisation. The distinction between sibilants and non-sibilants is borne out by parameters

such as relative/overall amplitude, duration, transition (locus equations/F2 onset) and frication noise information (see below) [5, 7, 10, 16]. Here, we will observe to what extent spectral properties of voiceless fricatives are affected by the sound transmission characteristics of different FCGs. We can only report on the main effects of the two independent within-subject variables *place of articulation (POA)* and *FCG*, with the intention of giving an overall impression of the data. The dependent variables under consideration are *peak* and the first four statistical moments of the FFT spectra (m1-m4), i.e. *centre of gravity*, *variance*, *skewness* and *kurtosis*. Statistical analysis was carried out using a repeated-measures ANOVA.

3. RESULTS

Figures 1-4 show the effects of the FCGs averaged across tokens, syllable positions and gender. There is a significant main effect at p<.01 of both POA (F(3,6)=19.41) and FCG (F(7,14)=6.24) on *centre of gravity*. Our data replicate the repeatedly reported finding that m1 distinguishes consistently between /s/ and /ʃ/ (see Figure 1). The higher m1 for /ʃ/ in the TAP condition may be the result of an absence of lip-rounding when speaking with tape across the mouth [6]. For some FCGs (especially RUB/HEL/TAP) we expect more sound absorption by the mask material, an assumption which was supported by intensity measures and auditory inspection. Under these guise conditions, m1 is lower for the non-sibilants, which may be more prone to damping of higher frequencies due to their greater spectral diffuseness and overall lower energy [6, 8, 15].

Figure 1: Centre of gravity (kHz) for all fricatives, averaged across tokens, syllable positions and gender.



The *variance*, i.e. the squared standard deviation of m1, is higher for the non-sibilants, which is again predictable given their diffuseness and lower intensity. Thus, we found a significant main effect of POA (F(3,6)=100.78, p<.01), but not of FCG (F(7,14)=1.04, p=.45).

The automatically detected *peak* values were manually corrected in order to eliminate tracking errors. These were caused by distortions in the signal which may, for instance, have been the result of rustling noises of certain mask materials or high-frequency whistling sounds caused by impedance of the airstream by the FCG. Figure 3 shows that all fricatives give rise to very high peaks in the F4 or higher range, and similar to m1, the peak distinguishes the sibilants, but less clearly so the highly variable non-sibilants. There is a significant effect of POA (F(3,6)=12.64, p<.01), but no main effect of FCG (F(7,14)=.65, p=.71).

Figure 2: Variance of m1 (MHz) for all fricatives, averaged across tokens, syllable positions and gender.

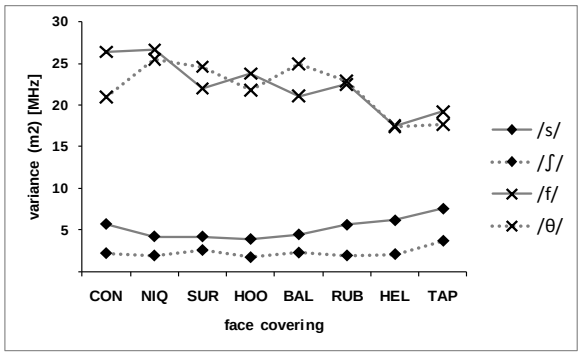


Figure 3: Spectral peak (kHz) for all fricatives, averaged across tokens, syllable positions and gender.

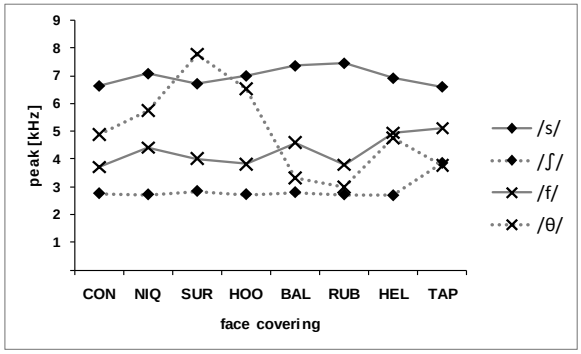
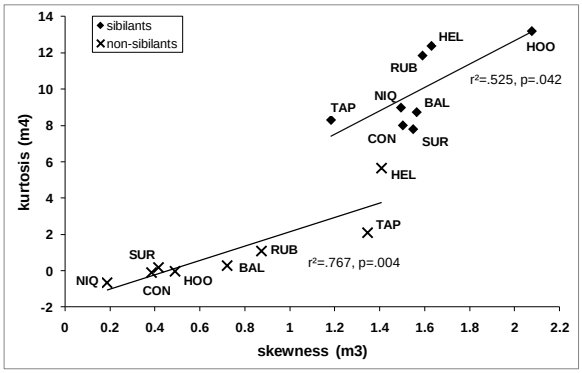


Figure 4: Skewness and kurtosis (dimensionless) for sibilants and non-sibilants, averaged across tokens, syllable positions and gender.



Reports in the literature for *skewness* and *kurtosis* are less consistent [5, 15]. In our data, both m3 and m4 were significant at p<.01 for both POA (F(3,6)=29.60; F(3,6)=17.42) and FCG (F(7,14)= 4.62; F(7,14)=6.47). Figure 4 shows a scatter plot for both classes of fricatives, with the means for m3 and m4 determining the data point for each FCG. There is a trend for the means to increase for most face masks compared to the control condition (CON), with a significant positive correlation found for the non-sibilants (r^2=.77, p<.01) and the sibilants (r^2=.53, p<.05). This tendency is dependent on the mask material, and the greatest effect can be observed for RUB, HEL and TAP.

4. CONCLUSION

This paper reports initial results from a spectral analysis of a subset of audio data from the 'face cover' corpus. Five parameters capturing the spectral properties of fricatives were demonstrated to be significantly affected by certain FCGs. These findings will be of particular interest in connection with results of future speech perception tests, in which participants will be confronted with the auditory consequences of the FCGs under varying A(V) conditions. The higher intensity of sibilants has previously been shown to distinguish them perceptually from non-sibilants. As a decrease in amplitude leads them to be confused with non-sibilants (not vice versa) [6], we speculate, for instance, that certain masks will attenuate the overall intensity of sibilants such that they will be increasingly confused with non-sibilants. In the course of the project, more speech material and phoneme classes will be included, and the forensic relevance – in particular the implications for ear-witness testimony, and for speech perception more generally – will be assessed.

5. REFERENCES

[1] BBfor2: Bayesian Biometrics for Forensics. *http://www.bbfor2.net*

[2] Coniam, D. 2005. The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-SARS study in Hong Kong. *Language Assessment Quarterly* 2, 235-261.

[3] Flipsen, P. Jr., Shriberg, L., Weismer, G., Karlsson, H., McSweeny, J. 1999. Acoustic characteristics of /s/ in adolescents. *JSLHR* 42, 663-677.

[4] Fuchs, S., Weirich, M., Kroos, C., Fecher, N., Pape, D., Koppetsch, S. 2010. Time for a shave? Does facial hair interfere with visual speech intelligibility? In Fuchs, S., Hoole, P., Mooshammer, C., Zygis, M. (eds.), *Between the Regular and the Particular in Speech and Language*, Frankfurt/M.: Peter Lang, 247-264.

[5] Haley, K.L., Seelinger, E., Mandulak, K.C., Zajac, D.J. 2010. Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach. *Journal of Phonetics* 38(4), 548-554.

[6] Harrington, J. 2010. Acoustic Phonetics. In Hardcastle, W.J., Laver J., Gibbon, F.E. (eds.), *The Handbook of Pho-netic Sciences* (2nd ed.). Oxford: Wiley-Blackwell, 81-129.

[7] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of English fricatives. *JASA* 108, 1252-63.

[8] Llamas, C., Harrison, P., Donnelly, D., Watt, D. 2009. Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistics* (Series 2) 9, 80-104.

[9] Lovitt, A., Allen, J.B. 2006. 50 years late: repeating Miller-Nicely 1955. *Interspeech* Pittsburgh, 2154-57.

[10] Maniwa, K., Jongman, A., Wade, T. 2009. Acoustic characteristics of clearly spoken English fricatives. *JASA* 125(6), 3962-73.

[11] Phatak, S.A., Lovitt, A., Allen, J.B. 2008. Consonant confusions in white noise, *JASA* 124(2), 1220-33.

[12] Preminger, J.E., Lin, H.-B., Payen, M., Levitt, H., 1998. Selective visual masking in speechreading. *JSLHR* 41, 564-575.

[13] Rosenblum, L.D., Saldaña, H.M. 1996. An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psy.: Human Perc. & Perf.* 22(2), 318-331.

[14] Schwartz, J.L., Berthommier, F., Savariaux, C. 2004. Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition* 93(2), 69-78.

[15] Shadle, C., Mair, S.J. 1996. Quantifying spectral characteristics of fricatives. *Interspeech* Philadelphia, 1521-1524.

[16] Tabain, M., Watson, C. 1996. Classification of fricatives. *Proc. 6th Aust. Int. Conf. Speech Sci. Technol.* Adelaide, 623-628.

[17] Watt, D., Llamas, C., Harrison, P. 2010. Differences in perceived sound quality between speech recordings filtered using transmission loss spectra of selected fabrics. Paper presented at the *2010 IAFPA Conference*, Trier.

[18] Zhang, C., Tan, T. 2008. Voice disguise and automatic speaker recognition, *Forensic Science International* 175(2-3), 118-122.

[1] In the present study, different materials from those in [8, 17] are used. In keeping with these studies we will continue referring to the face coverings as FCGs.

[2] One mask, a balaclava with a mouth hole, was excluded from the present study.

[3] /C_1ɑh/ and /ŋɑC_2/ syllables are phonotactically illegal in English, so /h/ and /ŋ/ appear only in onset and coda positions, respectively.