On whispered speech and its social functions

Marzena Żygis

Whispering in everyday communication

The phonation of the human voice is wide and varied, ranging from normal (modal) to whispered speech with many different speech modes in-between, including voices that are breathy, creaky, harsh or falsetto. While modal speech is undoubtedly the most frequently used, whispering is in fact a relatively common mode of communication, usually applied for the purpose of discreet communication. It plays a clearly positive role in the private domain, where it serves to signal bonding with the addressee. In this regard, Cirillo (2004) states that males use whispering significantly more often to express tenderness than females do. When used in public, whispering serves to (i) avoid disturbing someone (24%), (ii) communicate a secret message to a specific person (22%), (iii) confirm affiliation to the addressee (17%), (iv) initiate a playful encounter (14%), and (v) attract the attention of or induce curiosity in members of an audience (11%) (Cirillo 2004). It should, however, not be overlooked that whispering can also induce a sense of alienation or a negative feeling of being socially excluded.

Whispering also triggers the so-called Autonomous Sensory Meridian Response (ASMR), known as *scalp massage* or *neurological shortcut to relaxation*, in which "individuals experience a tingling, static-like sensation across the scalp, back of the neck and at times further areas in response to specific triggering audio and visual stimuli. This sensation is widely reported to be accompanied by feelings of relaxation and well-being" (Barratt & Davis 2015:1). Besides whispering, the ASMR can also be triggered by crisp sounds and slow movements. The phenomenon has also recently been reported in the press (Morgenpost 20.05.2019), where it was pointed out that some YouTube videos on "ASMR" channels had a huge number of clicks, with the most popular reaching 84 million. Neurologists warn that our knowledge about ASMR is extremely limited, stating that further studies testing brain activities should illuminate this already widely observed but as yet not well understood phenomenon.

Whereas most speakers sometimes voluntarily whisper in order to meet their various intentions, there are speakers for whom whispering remains the only available mode of spoken communication. The inability to use normally voiced speech can be either temporary – due, for instance, to throat inflammation – or permanent, due to, say, a tracheotomy. In addition, gradual loss of voice accompanies neuro-degenerative diseases such as Parkinson's (an estimated seven to ten million people worldwide suffer from Parkinson's disease:

https://parkinsonsnewstoday.com/parkinsons-disease-statistics/). Furthermore, gradual voice loss can also be caused by advancing age: older people often have problems with phonation even if not suffering from any particular health condition.

Articulation and acoustics of whispered speech

What exactly happens in terms of articulation and acoustics when we whisper? First of all, as transillumination studies have revealed, our vocal folds stop vibrating. In normal speech, the vocal folds also cease to vibrate when voiceless sounds are produced. In languages such as Berber, where syllables can be composed of exclusively voiceless sounds, voiceless words are not rare (Riduane et al. 2005). Several publications repeat the distinction that, whereas whispering is characterized by a triangular opening of the cartilaginous glottis (a rather small area), voicelessness is produced by a large glottal opening (see Laver 1994:190). Figure 1 illustrates the difference in glottal opening between breath (i.e. voiceless in Laver's terms) and whispered phonation. However, experimental studies do not always support this observation. For instance, Zeroual et al. (2005) have shown that the glottis can be larger in whispered speech. Moreover, glottal aperture differences observed in normal speech between voiceless consonants (e.g./t/) are neutralized in the production of whispered stops but tend to be maintained in the production of whispered fricatives i.e. the glottal opening is larger in voiceless whispered fricatives (e.g. /s, χ/) than in the corresponding whispered voiced fricatives /z, ʁ/).

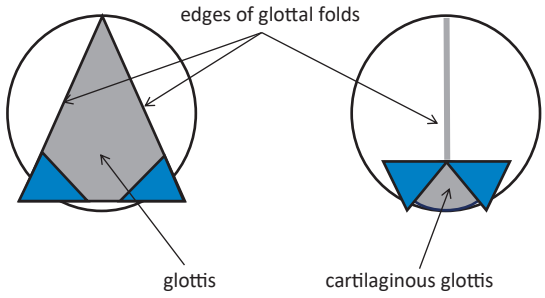

FIGURE 1

Schematic views of the glottis in breath (left) and whisper phonation (right) based on Laver (1994: 191)

The acoustic consequence of the lack of vocal fold vibration is the lack of fundamental frequency (F0).* In Figure 2 the oscillograms (upper part) and spectrograms (lower part) pre-

* F0 is defined as the number of vocal fold cycles per second. If F0 equals 200 Hz, it means that the vocal folds execute 200 cycles per second.

sent the same sentence in three different speech modes – i.e. modal (left), whispered (centre), and semi-whispered speech mode (right). The blue line in spectrograms indicates the fundamental frequency. While it is fully present in voiced speech and slightly interrupted in semi-whispered speech, it is entirely absent in whispered speech. In addition, we can also observe that the intensity of the sentences (corresponding to perceived loudness) differs as it is lower in whispered and in semi-whispered speech than in voiced speech, i.e. the amplitude of the wave in the oscillogram is lower in the last two cases than in modal speech.

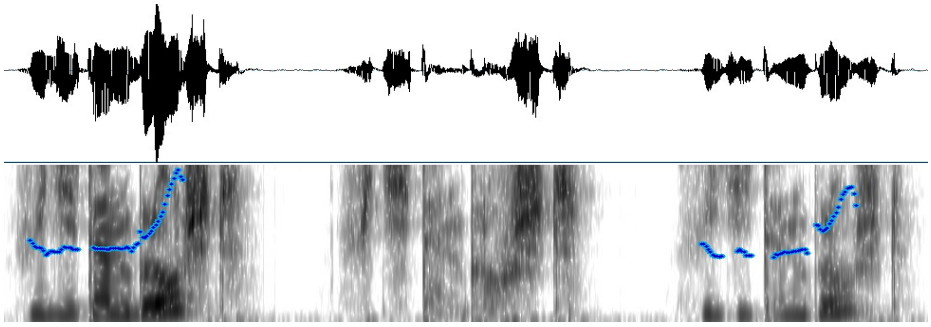

FIGURE 2

Spectrogram of the Polish sentence *Widzi ten bluszcz?* 'Does he see the ivy?' in modal (left), whispered (centre) and semi-whispered speech (right)

A prevailing number of studies investigating whispered speech focus on vowels (e.g. /a, e, i/). It has been shown that whispered vowels are characterised by decreased amplitudes, higher formants and longer duration compared to voiced vowels (see e.g. Ito, Takeda, & Itakura 2005). Very little attention has been devoted to consonants in whispered speech. This is most probably due to the difficulty of measuring them: whispered consonants do not only lack periodic glottal excitation, but due to the noise generated during their production, the spectrograms are also dominated by strong aperiodic energy which might posit challenges for researchers.

But what is even more challenging for researchers is the investigation of prosody, i.e. changes in fundamental frequency (F0), duration, intensity (and others), which are critical for speech communication because they help to identify the boundaries of words and sentences. By means of prosody, speakers are also able both to highlight the most important piece of information in a given sentence – by using linguistic focus – and to produce the intended sentence type, e.g. question or statement. Finally, individual characteristics of speakers' emotions, attitudes or intentions are also encoded in prosodic cues.

Despite the undeniable role of prosody in communication, research on voiced speech has concentrated (again) on vowels and sonorants (e.g. /l, r, j/), mainly investigating their parameters, especially F0 changing across syllables, words and sentences. The evidence concerning prosody in whispered speech is even more scant and limited to vowels with very few exceptions. In our study (Żygis et al. 2017), we were able to show that not only vowels but also consonants encode prosodic cues during whispering. Our data revealed that whispered utterance-final fricatives and affricates change their parameters as a function of intonation. In questions with a rising intonation, the sounds are produced with a higher intensity, a higher centre of gravity (spectral mean), higher-frequency spectral peaks and higher spectral standard deviation values than in statements where the intonation does not rise. Conversely, skewness and kurtosis (peakedness) of the spectrum are lower in questions than in statements. Importantly, some spectral features of sibilants, including spectral slopes, show more pronounced differences in the whispered than voiced speech mode. The finding that some cues are more pronounced in whispered speech lets us conclude that they may compensate for the absence of fundamental frequency in this mode. Such relations, in which one cue compensates for the absence or reduced occurrence of another, are known in the literature as *trading relations* and are related to acoustics and perception (e.g. Parker et al.1986).

Trading relations in whispered speech

Trading relations are lively discussed in terms of speech and gestures. According to the *trade-off* hypothesis, when speaking becomes more difficult, the likelihood of a head or hand gesture to 'take over' some of the communicative load is higher; and conversely, when gesturing becomes harder, speakers will rely more on speech (de Ruiter et al. 2012). The hypothesis has been tested by studying the interplay of speech and gestures in voiced speech. It remains, however, entirely unclear what happens to the acoustic signal and gestures when the former becomes degenerated, as is the case in whispered speech, and therefore harder to understand. Do gestures then convey prosodic details and therefore contribute more intensively to the understanding of speech? Does the hypothesis also apply to oro-facial expressions such as lip opening or eyebrow movements?

Questions of this type have thus far scarcely been addressed in the literature. Dohen & Loevenbruck's (2008) study, which revealed that perception of prosodic focus in French whispered speech is difficult to discern when it is based on acoustic signal, is the exception to this rule. The study shows that visual signals, such as lip opening produced while whispering, decidedly enhance perception of prosodic focus, leading to much faster reaction times. The results of our own study (Żygis et al. 2017) show that whispered vowels in German are produced with a higher lip aperture than normally produced vowels, compen-

sating, in our interpretation, for the lack of F0. The lip aperture is also higher in questions than in statements. We found that the right eyebrow is higher in questions than in statements, too, though the effect was limited to only a few speakers and was not speech-mode dependent.

Finally, as speech communication does not only happen when the interlocutors see each other, it would be interesting to see whether and how both speech and gesture change in function of visibility. In this regard, Cvejic et al. (2012) were able to show that the production and perception of prosodic focus and phrasing differs when interlocutors (i) only hear each other as opposed to (ii) when they hear and see each other. Acoustic differences were greater in the first condition. Likewise, the perception of narrow focus (i.e. the most important and therefore heavily stressed word in a sentence) and echoic question phrasing was better in the auditory-only condition. The results were interpreted again in terms of compensation effects for the lack of visual cues: speakers exaggerate the acoustic (prosodic) cues to compensate for the lack of complementary visual cues (head gestures, eyebrow movements etc.). This conclusion can lead to the hypothesis that the strongest and most pronounced cues are to be expected in whispered speech when the speakers do not see each other. In fact, our study (Żygis & Fuchs to appear) has confirmed this hypothesis. Our results show that speakers compensate for the fact that they cannot see each other, i.e. they try to be properly understood by using wider lip opening than when they are visible to each other. They open their lips even wider when they whisper (see Figure 3 for the experimental setting with an artificial wall between the speakers).

Figure 3

Experimental setting for an invisible mode with an artificial (dividing) wall between the speakers

Summary

Understanding whispered speech can help us to discern the underlying mechanisms of normal speech. One such mechanism is the trade-off relations found between (i) acoustic cues and (ii) speech signal and gestures. Compensation effects also appear when communication does not happen face-to-face but takes place when speakers cannot see each other.

Research on whispered speech is useful not only for understanding speech and communication as such, but can also be applied in many areas of everyday life. If we understood the nature of whispered speech better, we could develop algorithms converting whispered speech to voiced speech. Such algorithms would help relieve the daily challenges of the ageing and clinical populations suffering from ailments such as throat cancers or neurodegenerative diseases including Parkinson's. They could be used in telecommunications, with whispered speech converted into voiced speech, as well as in security systems and forensics (see Figure 4).

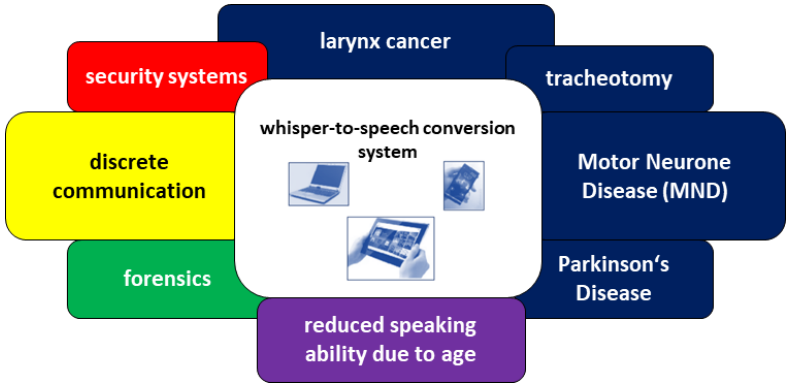

FIGURE 4
Possible applications of whispered speech

Despite great progress in the development of different types of devices that help to understand whispered speech, including (i) reconstruction of continuous voiced speech from whispered speech (McLoughlin et al. 2013) and (ii) esophageal speech enhancement (Caeiros & Meana 2010), these interfaces struggle when it comes to properly modelling the prosody of the spoken sentence. Against this background, naturalness and correct interpretation of speaker intentions are difficult to achieve. To overcome this difficulty, more research is called for in order to implement findings revealed by experimental acoustic and gestural analyses of both whispered and voiced speech.

Barratt, E. L. & N. J. Davis (2015). *Autonomous Sensory Meridian Response (ASMR): a flow-like mental state*. PeerJ the Journal of Life and Environmental Sciences 3, https://www.ncbi.nlm.nih.gov/pubmed/25834771.

Caeiros, M. A. V. & H. M. P. Meana (2012). Esophageal speech enhancement using a feature extraction method based on wavelet transform. In: Ramakrishnan S. (ed.) *Modern Speech Recognition Approaches with Case Studies*. TechOpen, DOI: 10.5772/49943

Cirillo, J. (2004). Communication by unvoiced speech: the role of whispering. *Anais da Academia Brasileira de Ciências* 76, 1-11.

Cvejic E., J. Kim & C. Davis (2012). Effects of seeing the interlocutor on the production of prosodic contrasts. Journal of the Acoustical Society of America 131, 1011-1014.

De Ruiter, J. P., A. Bangerter & P. Dings (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science* 4, 232–248.

Dohen, M. & H. Loevenbruck (2008). Audiovisual perception of prosodic contrastive focus in whispered French. *Journal of the Acoustical Society of America* 123, 3460-3460.

Ito, T., K. Takeda & F. Itakura (2005). Analysis and recognition of whispered speech. *Speech Communication* 45, 139-152.

Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.

McLoughin I., L. Jingjie & S. Yan (2013). Reconstruction of continuous voiced speech from whispers. *Proceedings of Interspeech*, Lyon, Fance, 1022-1026.

Parker, E. M., R. L. Diehl & K. R. Kluender (1986). Trading relations in speech and nonspeech. *Perception and Psychophysics* 39, 129-42.

Ridouane, R., S. Fuchs & P. Hoole (2005). Laryngeal adjustments in the production of voiceless obstruent clusters in Berber. In Harrington, J. & M. Tabain, *Speech Production: Models, Phonetic Processes and Techniques*, 275-297. New York: Psychology Press.

Zeroual, Ch., J. Estling & L. Crevier-Buchman (2005). Physiological study of whispered speech in Moroccan Arabic. *Proceedings of Interspeech* 1069-1072. Lisbon, Portugal.

Żygis, M., S. Fuchs & K. Stoltmann (2017). Orofacial expressions in German questions and statements in voiced and whispered speech. *Journal of Multimodal Communication Studies 4, 87-92.*

Żygis, M., D. Pape, L. Koenig, M. Jaskula & L. Jesus (2017). Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. *Journal of Phonetics 63, 53-74*. http://www.sciencedirect.com/science/article/pii/S0095447017300645

Żygis, M. & S. Fuchs (to appear). How prosody, speech mode and speaker's visibiity influence lip aperture. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia.*