THE VISUAL PROMINENCE OF WHISPERED SPEECH IN SWEDISH

Zofia Malisz, Patrik Jonell, Jonas Beskow

KTH Royal Institute of Technology, Stockholm, Sweden
[malisz, pjjonell, beskow]@kth.se

ABSTRACT

This study presents a database of controlled speech material as well as spontaneous Swedish conversation produced in modal and whispered voice. The database includes facial expression and head movement features tracked by a non-invasive and unobtrusive system. We analyse differences between the voice conditions in the visual domain paying particular attention to realisations of prosodic structure, namely, prominence patterns. Analysis results show that prominent vowels in whisper are expressed with a statistically significantly a) larger jaw opening, b) stronger lip rounding and protrusion, c) higher eyebrow raising and d) higher pitch angle velocity of the head, relative to modal speech.

Keywords: prosody, voice quality, whisper, head gestures, orofacial gestures

1. INTRODUCTION

Prosodic variation is expressed multimodally [28]. The term audiovisual prosody originates in the widely studied interaction of orofacial gestures that correlate with prosodic structure [23]. Similarly, the links between modal speech and speech-accompanying gesture in the expression of prosody are especially close. Especially prosodic prominence is often produced multimodally [22, 1, 3]. The gestural enhancement of prosodic prominence serves interactive functions, e.g. by showing a higher degree of attention in dialogue or to express semantic functions, e.g. by signaling information focus [5]. [22] found that prosodic prominences accompanied with "visual beats" produced by head movement showed interactions with specific acoustic exponents of prominence, namely changes in the formant structure and duration patterns, when compared to prosodic prominences produced without such visual beats. Moreover, visual signals that are not directly related to speech articulation, such as head movement, support the perception of prosody (similarly [6]) and facilitate comprehension [27].

However, the precise nature of the interaction between visual prosody and acoustically expressed prosody, e.g. in prominence highlighting, is not very well understood [3]. One important question is whether modalities parallel or complement each other. For example, is a lack in verbally produced prominence balanced by gesture or do the various modalities contribute to prominence additively? Recent studies [11] found evidence for an additive effect of both modalities in prominence highlighting. [7] investigated whether production of prosodic focus and phrasing contrasts was modified in auditory and face-to-face setting by conducting acoustic and perceptual studies. Acoustically realised narrow focus and phrasing contrasts were greater in the audio-only setting than in the face-to-face setting, indicating that gesture adds to and modulates acoustic prosody.

Most of the literature on the relationship between prosody and gesture is based on modal speech. Similarly, the studies that exist on the acoustic prosody of whispered speech, do not include the visual dimension. These acoustic and perceptual studies show that F0 is absent in whisper but intonation is still discernible [13, 17]. [12, 14, 15] studied the perceptual discrimination of prosody in declarative and interrogative sentences in French. Results showed that discrimination was based on different acoustic cues in modal vs. whispered speech. The high-frequency region was the main cue in whisper, whereas in modal speech, the listeners relied mainly on the low-frequency region. Additionally, perception of prosody was based on spectral rather than temporal auditory cues. Dutch speakers in turn, used mainly secondary cues to intonation in expressing boundary tones in whisper, that is, they actually exhibited minimal compensatory cues to intonation.

As indicated, we find a very limited number of studies focusing on the audio-visual prosody of whisper [8, 9]. Questions on the interplay of acoustic and visual structures in the expression of prosody motivated the studies by [8, 9]. The authors were interested in the multimodal expression of focus and used whispered speech mode as a paradigm to elicit different levels of visual focus. The results revealed that an increase in lip area and jaw opening correlate with contrastive focus in French. Postfocal syllables on the other hand, showed a significant reduction of lip area and jaw opening. These two patterns correspond to hyper- and hypo-articulation visible in the lip movements.

The only results on audiovisual features of whisper in Swedish that we are aware of is [2]. The analysis was conducted for the purposes of creating an animated agent. The studied participant read sentences and words while two meters away from a listener in modal voice in quiet (baseline condition), modal voice in noise (Lombard speech) and in whispered speech. The participant was asked to make sure that he is understood by the listener. The results showed that the global inter-lip velocity was highest in whisper overall and twice as high compared to the baseline, suggesting hyper-articulation effects.

[7] conducted a study of audiovisual prosody in modal speech. Their results suggested that in the audio-only setting, speakers exaggerated acoustic cues to prosody to compensate for the lack of complementary visual cues (head gestures, eyebrows etc.).

In the present study, we similarly hypothesise that mutual compensation effects across modalities and speech modes will be stronger in case of inherently degraded whisper than in modal speech. We analyse prominent vowels in a reading task in Swedish. We explore the impact of whisper vs. modal speech modes on several visual correlates of prominence as indicated, or similar to those indicated, in the studies discussed above: jaw opening, lip rounding and protrusion, eyebrow raising and the velocity of head movement on the pitch (vertical) axis.

2. THE DATABASE

2.1. Data recording and tracking

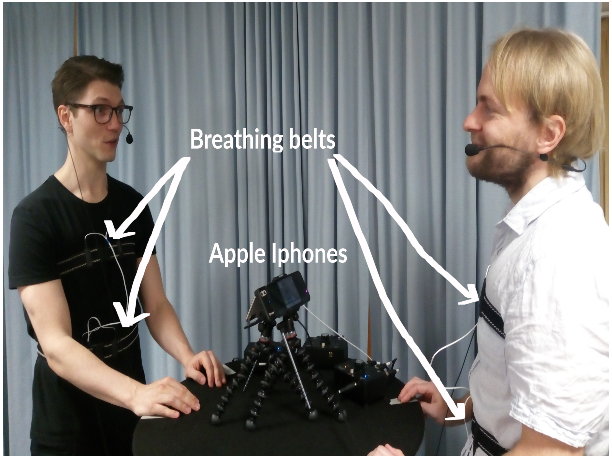
Figure 1 presents the database recording setup. The speech signal was captured using high-quality close-talking microphones on separate channels. Respiration was tracked by displacement of the chest and abdomen, captured by elastic belts worn around the thorax (Respiratory Inductance Plethysmography, RIP) [29]. Visual signals, i.e., orofacial gestures and head movement, were captured using two Apple iPhones (XS 2018 models, with 12 Mpx cameras plus a depth camera), each mounted facing one of the participants. For purposes not within the scope of the present analysis, we also used a Forward Looking Infrared Camera (FLIR) for thermal imaging of each participant in this setup.

The visual data was extracted via a custom application developed using the Apple Software Development Kit (iOS 11.0+) and the AppleAR framework [18]. The feature set in AppleAR is very extensive and was created in order to allow independent developers to create their own virtual characters based on real-person tracking data. The face tracking configuration detects the participant's face in the view of the iPhone's True Depth 3D camera and represents defined facial features describing facial expressions. We tracked over 50 features using this framework: 14 for the eyes, 26 for the jaw and mouth plus 10 for eyebrows, cheek and nose. We also extracted a vector defining the movement of the head in three dimensions equivalent to turning the head right or left, up or down and forwards or backwards.

Figure 2 presents what the orofacial feature vector represents in AppleAR. If the coefficient of a feature is zero, the tracker detects a neutral shape and no movement, while the coefficient of 1.0 defines maximum movement and extent of the gesture. Typically, the system returns a value between these two extremes in 30 frames per second.

Signals were synchronised using FARMI [19] - a framework that allows for flexible and modular collection of data from multiple sensors. Data streams and messages are synchronised by off-setting timestamps using a delta function from a central time server, ensuring that timestamps for each data module are comparable.

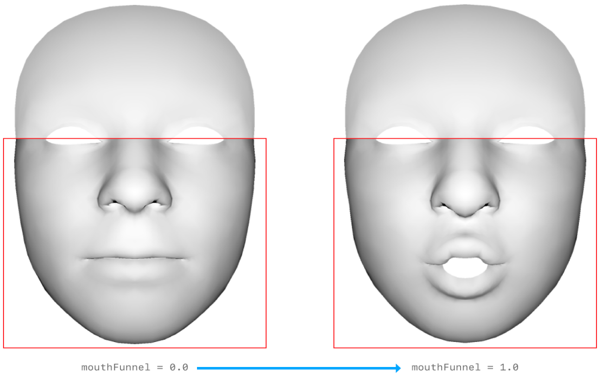
Figure 1: Studio setup



2.2. Experimental conditions

Each participant was recorded separately for the first two reading tasks: a) they read a one-page "easy Swedish" text about the education system in Sweden and b) they responded to questions asked by the experimenter. The questions were built to elicit broad and narrow focus (following the experimental design and subset of stimuli in [16]). Task c), the last task in the database, involved participant pairs who were asked to converse with each other for approx. 5-7 minutes on a topic of their choice. All tasks were varied by voice mode: modal and whis-

Figure 2: A feature tracked by the AppleAR framework illustrated in an ARFaceGeometry. In this case, a lip rounding and protrusion gesture into and open shape” (MouthFunnel). Two states are shown with coefficient at value = 0.0 (neutral shape, left), and at value = 1.0 (maximum movement, right).



pered, the order of performing in each voice mode was counterbalanced.

Regarding some possible proximity or communicative situation factors, in task a), the participants were not asked to articulate in whisper or modal voice over distance towards an interlocutor as in [8] or [2]. They were requested to speak normally, as would be natural in a reading situation. In b), the elicitation method involved a repetitive and predictable form of read dialogue with the experimenter. In comparison, task c) involved the most ecologically valid, spontaneous form of interaction from a proximity shown in Figure 1.

2.3. Speakers

We recorded 10 participants in 5 sessions so far in this setup. Participants were native speakers of Swedish (incl. 1 bilingual, Swedish-German speaker) from different parts of Sweden, all resided in Stockholm for at least several years. They had no known hearing or speaking impairments.

3. THE PRESENT ANALYSIS

The following analysis includes data from 6 participants on 561 observations in task a): reading of a text in Swedish in the two speech modes. We analysed prominent vowels in about a minute of speech in each mode per participant in this task.

3.1. Speech and multimodal data processing

Whispered segments were annotated manually by a phonetic expert in Praat. Modal speech was force aligned using WebMaus [20] for Swedish and subsequently inspected in Praat for corrections and consistency in comparison to the whispered speech segmentation. Prominent vowels were indexed in both modes using linguistic criteria, i.e., candidates were first marked in the read text according to lexical stress and phrase accent rules in Swedish, they were then perceptually evaluated if they were actually prominent also in realisation (by the first author and a native speaker of Swedish).

The FARMI recording architecture ensures that all signals are captured in synchrony. We additionally verified accuracy and synchrony by visualising vectors with tracking data for each studied feature along with segmentations for each speech mode in ELAN and compared them to the video. We established that all signal streams were synchronised.

We extracted maxima of the orofacial feature vectors for each vowel marked as prominent in both speech modes. In the present analysis, we used orofacial features that describe the degree of jaw opening (ARKit: JawOpen), lip rounding with protrusion (MouthFunnel) and the raising of both eyebrows (BrowInnerUp).

A 4x4 head movement transformation matrix is put out by ARKit for each frame. We processed the head data using previous Python tool implementation for multimodal data [21]. The values in the matrix were converted to Euler angles and smoothed with Savitzky-Golay filtering (window length = 7). We extracted the pitch radial velocity (vertical head movement speed) that the filter produces and recorded the maximum absolute value within frames co-occurring with each analysed vowel.

3.2. Statistical analysis methods

The visual and speech data were integrated and analysed in R. Separate linear mixed models with each orofacial modality (jaw opening, lip rounding, brows up) and the head pitch velocity vector as the response were formulated.

Random structure was maximised and evaluated using likelihood ratio-based model comparisons. Random slopes resulting either in convergence issues (after 100k iterations) or perfect correlations were removed. P-values were estimated via the Satterthwaite approximation with lmerTest [24].

We entered Mode (modal, whisper) and Vowel Quality as predictors as well as Word and Speaker as random intercepts. Interaction of Mode and Vowel was tested in each model, however, no significant interactions were found.

4. RESULTS

4.1. Jaw opening

The whisper mode had a significant positive effect on jaw opening (est = 1.4, p<.001) relative to modal speech. The open-mid vowel /ɛː/ (est = 6.2, p<.01) also increased jaw opening relative to the vocalic grand mean.

4.2. Lip rounding and protrusion

Lips rounded and protruded into an open shape more in whispered speech (est = 0.02, p<.001) than in modal speech. We also observed several significant effects of vowels expected from their qualities: rounded or open vowels /ø, ɑː, ɔ, oː, ʊ/ exhibited positive effects and front unrounded vowels /ɛ, ɛː, ɪ, eː, iː/ showed negative effects, all remaining ones did not significantly deviate from the vocalic grand mean.

4.3. Eyebrow raising

Speaking in whisper significantly increased eyebrow raising on prominent vowels (est = 1.4, p<.001), so did the vowel qualities /ɛː/ (est = 1.2 , p<.001) and /oː/ (est = 1.1, p<.05).

4.4. Head movement

In terms of effects on vertical head movement velocity, the results suggest that head movement reaches higher speed in whispered prominent vowels (est = 4.4, p<.05) than in modal prominent vowels in this task.

5. DISCUSSION

We presented a database with several controlled and conversational conditions built to study the differences between whispered and modal speech in the audiovisual domain. We also presented a non-invasive tracking method of orofacial and head movements in which most tracking components are available commercially with an easily buildable and flexible synchronisation of breathing, speech and visual features such as head movement and orofacial expressions.

An analysis of Swedish prominent vowels was conducted in one of the database tasks. Focus was put on the impact of whispered vs. modal voice in the orofacial and head gesture domain.

First of all, the results suggest effects consistent with [10] for French and [2] for Swedish. The magnitude of the gestures is larger in whispered speech than in modal voice in a task that involves reading a running text aloud in the two modes.

Specifically, in Swedish connected speech while reading in whisper, the degree of jaw opening is larger. Similarly, lip protrusion into an open shape (a gesture presented in Figure 2) is also more pronounced in prominent whispered vowels than prominent modal vowels. Eyebrow raising, a feature associated largely with prosodic emphasis in audiovisual prosody studies[3] is positively affected by whisper: eyebrows are raised higher when whispering prominent vowels than modal prominent vowels.

We hypothesised that an augmentation of orofacial and head movement cues should occur in the inherently degraded mode of whisper, as speakers try to compensate for channel deficiencies in order to communicate effectively. Such a heightened distinctiveness is consistent with information-theoretic accounts of communication [4] as it increases the redundancy of the deficient signal [26] (a related perspective is the Hyper-&Hypoarticulation theory [25]).

It should be noted that our analysis was based on reading aloud in the two modes while alone in the studio booth, i.e. without elicitation of a particularly clear articulation or an explicit addressee present, as in [10, 2]. However, presumably, as reading aloud assumes an implicit addressee, we still do find listener-oriented, redundancy effects in the visual dimension.

6. CONCLUSIONS AND FUTURE WORK

We conclude that our analysis indicates that, in information-theoretic terms, there is compensation for the degraded qualities of the whispered speech channel in a reading aloud task. However, we see the necessity to carefully consider all articulatory and prosodic influences and constraints that might interact with the channel effect. This future study goal will be explored in an analysis of task b) in the database: a focus- and accent-differentiated whispered and modal data and in task c) involving spontaneous conversation in both modes.

7. ACKNOWLEDGEMENTS

We thank Christina Lindahl for help with annotation of whispered speech data and Simon Alexanderson for discussions. ZM is supported by the Swedish Research Council grant no. 2017-02861 “Multimodal encoding of prosodic prominence in voiced and whispered speech”. PJ and JB are supported by the Swedish Foundation for Strategic Research Grant No.: RIT15-0107.

8. REFERENCES

[1] Al Moubayed, S., Ananthakrishnan, G., Enflo, L. 2010. Automatic prominence classification in Swedish. *Speech Prosody 2010, Workshop on Prosodic Prominence, Chicago, USA*.

[2] Alexanderson, S., Beskow, J. 2014. Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech & Language* 28(2), 607–618.

[3] Ambrazaitis, G., House, D. 2017. Multimodal prominences: exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication* 95, 100–113.

[4] Aylett, M., Turk, A. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47, 31–56.

[5] Beskow, J., Granström, B., House, D. 2006. Visual correlates to prominence in several expressive modes. *INTERSPEECH 2006* Pittsburgh. 1272–1275.

[6] Cvejic, E., Kim, J., Davis, C. 2010. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication* 52(6), 555–564.

[7] Cvejic, E., Kim, J., Davis, C. 2012. Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition* 122(3), 442–453.

[8] Dohen, M., Loevenbruck, H. 2008. Audiovisual perception of prosodic contrastive focus in whispered French. *The Journal of the Acoustical Society of America* 123(5), 3460–3460.

[9] Dohen, M., Lœvenbruck, H. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech* 52(2-3), 177–206.

[10] Dohen, M., Loevenbruck, H., Callan, A., Callan, D., Baciu, M., Pichat, C., Harold, H. 2010. Multimodal perception of whispered and voiced prosody in french: A preliminary fmri study. *17th Annual Cognitive Neuroscience Society Meeting (CNS 2010)* 112–113.

[11] Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., Alías, F. 2014. Gesture synthesis adapted to speech emphasis. *Speech Communication* 57, 331–350.

[12] Heeren, W., van Heuven, V. 2014. The interaction of lexical and phrasal prosody in whispered speech. *The Journal of the Acoustical Society of America* 136(6), 3272–3289.

[13] Heeren, W., van Heuven, V. J., others, 2009. Perception and production of boundary tones in whispered Dutch. *INTERSPEECH 2009* Brighton, UK. 2411–2414.

[14] Heeren, W. F., Lorenzi, C. 2014. Perception of prosody in normal and whispered French. *The Journal of the Acoustical Society of America* 135(4), 2026–2040.

[15] Heeren, W. F. L., Lorenzi, C. 2014. Perception of prosody in normal and whispered French. *The Journal of the Acoustical Society of America* 135(4), 2026–2040.

[16] Heldner, M. 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics* 31(1), 39–62.

[17] Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H. 1996. Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study. *Journal of Voice* 10(2), 155–158.

[18] Inc., A. 2018. AppleARKit documentation. *https://developer.apple.com/documentation/arkit*.

[19] Jonell, P., Bystedt, M., Fallgren, P., Kontogiorgos, D., Lopes, J., Malisz, Z., Mascarenhas, S., Oertel, C., Raveh, E., Shore, T. 2018. FARMI: A framework for recording multi-modal interactions. *Proceedings of LREC* Miyazaki, Japan. 3969–3974.

[20] Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal processing via web services: The use case WebMAUS. *Proceedings of the Workshop on Service-oriented Architectures for the Humanities: Solutions and Impacts* Hamburg, Germany. 30–34.

[21] Kousidis, S., Malisz, Z., Wagner, P., Schlangen, D. 2013. Exploring annotation of head gesture forms in spontaneous human interaction. Proceedings of the Tilburg Gesture Meeting (TiGeR 2013).

[22] Krahmer, E., Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* 57(3), 396–414.

[23] Krahmer, E., Swerts, M. 2009. Audiovisual prosody-introduction to the special issue. *Language and speech* 52(2-3), 129–133.

[24] Kuznetsova, A., Bruun Brockhoff, P., Haubo Bojesen Christensen, R. 2016. *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-33.

[25] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H & H theory. In: Hardcastle, W. J., Marchal, A., (eds), *Speech Production and Speech Modelling*. Dordrecht: Kluwer 403–439.

[26] Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., Andreeva, B. 2018. Dimensions of segmental variability: interaction of prosody and surprisal in six languages. *Frontiers in Communication* 3, 25.

[27] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., Vatikiotis-Bateson, E. 2004. Visual prosody and speech intelligibility. *Psychological Science* 15(2), 133–137.

[28] Wagner, P., Malisz, Z., Kopp, S. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57, 209–232.

[29] Włodarczak, M., Heldner, M. 2016. Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking. *INTERSPEECH 2016* San Francisco, USA. 510–514.