AUTOMATIC PREDICTION OF INTONATION FROM SPEECH GESTURES

Application to voice substitution

Context

In oral interactions, speech prosody, which includes intonation, rhythm and timbre (voice quality), is a dedicated channel of communication that carries both speech structuring and expressive information. Yet, a growing number of speech pathologies (e.g., throat or neck cancer, etc.) affecting vocal folds vibration deprive patients of their control of intonation, thus severely impacting their speech intelligibility and social interactions [Morris et al., 2016]. Voice substitution solutions consist in reconstructing the degraded parts of a speech waveform from alternative sources of information [Schultz et al., 2017]. In the particular case of laryngeal impairment, a central aspect of voice substitution is the prediction of intonation from other speech production channels. In particular strong correlation was observed between prosodic variations (intonation, in particular), and speech co-occuring gestures such as movements of the lips [Dohen et al., 2004], the tongue [Krivokapić et al., 2017], the eyebrows [Cave et al., 1996], or the head [Wagner et al., 2014].

Objectives

Given these considerations, the goal of this internship is to build a system for ***the automatic prediction of intonation from orofacial gestures***, that will be integrated in the speech reconstruction system displayed in Fig. 1. The speech generation module is a whisper-to-speech conversion system available in the lab [Perrotin and McLoughlin, 2020] and the internship will focus on the orange blocks, according to the three following steps:

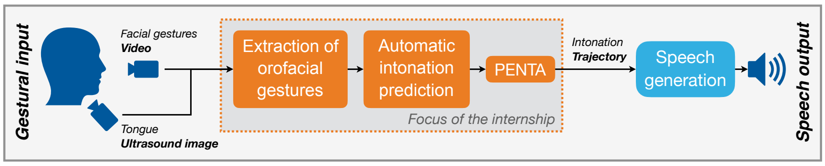

FIGURE 1. Full pipeline of the speech reconstruction system

(1) **Data acquisition:** We will record a corpus that includes series of utterances whose meaning differ only due to their intonation contours. For this sake, we will use strategies to elicit focalisation [Dohen and Lœvenbruck, 2009] and delimitative cues [Welby and Niebuhr, 2016] in sentences uttered by various speakers, and that are addressed to a listener in face-to-face interaction [Garnier et al., 2010]. We will measure facial movements with a camera placed in front of the participants' faces. Key regions of interest such as the lips, eyebrows and head movements will then be extracted with the MediaPipe or OpenFace toolkits. Tongue movements will be measured by ultrasound imaging that simply requires to place a probe under the chin, and extraction of features will follow [Hueber et al., 2010].

(2) **Corpus annotation and training of the PENTA model:** To facilitate the prediction of intonation, we will not directly predict the full intonation trajectory from gestures, but a ***prosodic score***, which is a sequence of descriptive labels for each intonation function (e.g., stressed/unstressed ; pre-, on, or post- position of focus; type of modality (question or assertion), etc.), annotated for each syllable of an utterance. The PENTA model [Xu and Prom-on, 2014] is a data-driven method for converting such prosodic scores into smooth intonation curves. This second step will consist in annotating the corpus with the prosodic score, and in training a PENTA model to generate of intonation curves from the prosodic score.

(3) **Automatic intonation prediction:** Finally, we will implement and compare several methods for the automatic prediction of a prosodic score from orofacial gestures that will be fed to PENTA. Linear to non-linear deep learning-based methods will be explored. In particular, we will quantify the context needed in the orofacial gestures for an accurate prediction, by exploring time-dependant architectures (dilated convolution [Wang et al., 2018], self-attention mechanisms [Chen et al., 2021]).

Tasks

The tasks expected during this internship are:

- Recording a multimodal (audio and gestures) pilot dataset of speech in interaction
- Training the PENTA model for the analysis and synthesis of intonation contours on our pilot dataset
- Developing first solutions for the automatic prediction of PENTA labels from speech gestures

Required skills

- Machine learning and signal processing.
- Speech science and technologies.
- Knowledge of Python is required for implementation.
- Strong motivation for dataset recording, methodology and experimentation.

Allowance

The internship allowance is fixed by ministerial decree (about 570 euros / month).

Contact

This internship will take place at GIPSA-lab, in the CRISSP and PCMD teams. It will be supervised by:

• Olivier PERROTIN	olivier.perrotin@grenoble-inp.fr	+33 4 76 57 45 36
• Marion DOHEN	marion.dohen@grenoble-inp.fr	
• Maëva GARNIER	maeva.garnier@grenoble-inp.fr	
• Thomas HUEBER	thomas.hueber@grenoble-inp.fr	

References

[Cave et al., 1996] Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and fo variations. In *International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 2175–2178. IEEE.

[Chen et al., 2021] Chen, X., Wu, Y., Wang, Z., Liu, S., and Li, J. (2021). Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ICASSP '21, pages 5904–5908, Toronto, Canada. IEEE.

[Dohen and Lœvenbruck, 2009] Dohen, M. and Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, 52(2-3):177–206. PMID: 19624029.

[Dohen et al., 2004] Dohen, M., Lœvenbruck, H., Cathiard, M.-A., and Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant french speech. *Speech Communication*, 44(1):155–172. Special Issue on Audio Visual speech processing.

[Garnier et al., 2010] Garnier, M., Henrich, N., and Dubois, D. (2010). Influence of sound immersion and communicative interaction on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 53(3):588–608.

[Hueber et al., 2010] Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288–300.

[Krivokapić et al., 2017] Krivokapić, J., Tiede, M. K., and Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, 8(3):1–26.

[Morris et al., 2016] Morris, M. A., Meier, S. K., Griffin, J. M., Branda, M. E., and Phelan, S. M. (2016). Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disability and Health Journal*, 9(1):140–144.

[Perrotin and McLoughlin, 2020] Perrotin, O. and McLoughlin, I. V. (2020). Glottal flow synthesis for whisper-to-speech conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:889–900.

[Schultz et al., 2017] Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., and Brumberg, J. S. (2017). Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271.

[Wagner et al., 2014] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57(Supplement C):209–232.

[Wang et al., 2018] Wang, X., Takaki, S., and Yamagishi, J. (2018). Autoregressive neural f0 model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1406–1419.

[Welby and Niebuhr, 2016] Welby, P. and Niebuhr, O. (2016). The influence of F0 discontinuity on intonational cues to word segmentation: A preliminary investigation. In *Speech Prosody*, pages 40–44, Boston, MA, USA. ISCA.

[Xu and Prom-on, 2014] Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication*, 57:181–208.