AUTOMATIC PREDICTION OF INTONATION FROM SPEECH GESTURES

Application to voice substitution

Apply here: https://emploi.cnrs.fr/Offres/Doctorant/UMR5216-CHRROM-033/Default.aspx

Context

In oral interactions, speech prosody, which includes intonation and rhythm, is a dedicated channel of communication that carries both speech structuring information (e.g., syntactic boundaries, contrastive focus) and expressiveness (e.g., attitudes or emotions). 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, we propose in this PhD thesis to: conceive a system for ***the automatic prediction of intonation from orofacial gestures*** that will be integrated in a real-time speech reconstruction system ; and evaluate this system ***in face-to-face interaction setups***. In particular, the three following steps will be addressed:

(1) **Data acquisition:** We aim at designing several interaction scenarios which will be used to both train and evaluate the automatic prediction of intonation system. In particular, *these scenarios must induce the production of sentences whose meaning only differ due to their intonation contours* (e.g., strategies to elicit focalisation [Dohen and Lœvenbruck, 2009] and delimitative cues [Welby and Niebuhr, 2016]). We will record a multi-speaker corpus in face-to-face interaction [Garnier et al., 2010], while measuring orofacial movements (e.g., with a camera and/or ultrasound imaging [Hueber et al., 2010]).

(2) **Automatic intonation prediction:** The PhD candidate will implement and compare methods for the automatic prediction of intonation from orofacial gestures. *The main challenge is to find the right balance between the use of the most recent deep learning-based methods and architectures that are fit for online processing*. For example, we may quantify the context needed in the orofacial gestures input for accurate prediction of intonation, by comparing time-dependant architectures (e.g., dilated convolution [Wang et al., 2018], causal / non-causal self-attention mechanisms [Chen et al., 2021], auto-regression [Wang et al., 2018]). To ease the prediction of intonation, the PENTA superpositional model of intonation [Xu and Prom-on, 2014] allows to tackle the prediction of each function of intonation (focus, delimitation, modality) independently.

(3) **Evaluation in a behavioural study:** The integration of the automatic prediction of intonation module in a real-time speech reconstruction system allows the user to get direct feedback of the quality of intonation prediction from the comprehension of the addressee. More interestingly, *the user may spontaneously adapt his/her input gestures to the system, to improve comprehensibility during the interaction*. The question of user adaptation to the system, and system adaptation to the user by fine-tuning the latter on new data will be explored in this last part, based on the various interaction scenarios designed earlier.

Environment and required skills

This PhD position is part of the SilentPitch ANR project (https://anr.fr/Projet-ANR-23-CE33-0016) which involves a puri-disciplinary team of researchers including machine learning, speech science, cognition and behavioural studies. If all disciplines are addressed in this PhD position, candidates without expertise in some of the areas listed below are nevertheless encouraged to apply.

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

A few-month visit to University College London to work with Prof. Yi Xu on the PENTA model is planned during the PhD. Also, we target an attendance to at least one international conference per year.

Grenoble Image Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE

Salary

Salary before tax is about 2135 euros / month.

Contact

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

- Olivier PERROTIN olivier.perrotin@grenoble-inp.fr +33 4 76 57 45 36
- Thomas HUEBER
- Maëva GARNIER
- Marion DOHEN

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.

[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.