Disentangling the Dimensions of Phonetic Variation: First Steps towards an Explanatory and Exploratory Research Tool in Phonetics

Petra Wagner[1], *Reinhold Haeb-Umbach*[2]
[1] *Faculty of Linguistics and Literary Studies, Bielefeld University, Germany*
[2] *Faculty for Computer Science, Electrical Engineering and Mathematics, Paderborn University, Germany*
petra.wagner@uni-bielefeld.de, haeb@nt.uni-paderborn.de

Abstract

In this paper, we present first evidence for a potential application of novel speech technological methods as a valuable tool for basic phonetics research. We describe a research program aiming at identifying the complex phonetic realizations underlying various dimensions of phonetic variation. This will be addressed with the help of recent approaches in unsupervised voice conversion and waveform generation. Concretely, we present a model for disentangling speakers' voice qualities and their linguistic-phonetic content, which can then be used to perform voice conversion across different dimensions of phonetic variation. The resulting signals are then "audible versions" of the phonetic dimensions of interest, and lend themselves to straightforward phonetic interpretation.

Introduction

Phonetics data are naturally "messy", as they are influenced by a myriad of dimensions, and typically more than those dimensions the researcher is primarily interested in (Pierrehumbert, 2004). Consequently, a researcher working on, e.g., dialectal variation will typically be unable to fully control for influential factors like gender, socioeconomic status, mood, or the size and shape of the vocal tract in her or his data. Beyond para-, socio- and extralinguistic variation, speech is naturally influenced by linguistic content, and slight changes may have tremendous influences on our phonetic variables of interest.

Hence, our data is often normalized, and many of our analyses are built on highly controlled settings, within homogenous groups of participants. This is of course highly problematic if the focus of investigation lies on spontaneous speech data, or if the phenomenon of interest is restricted to less controlled settings such as phonetic alignment, nonverbal vocalizations, turn taking phenomena or certain styles of speech that exclusively occur in less controlled environments (Wagner et al., 2015). In those cases, the target conflict between the need for control and the need for a lack of control cannot be solved straightforwardly. Hence, in state-of-the-art phonetics research, we need to devise novel, suitable methods for dealing with speech "in the wild".

Another issue of phonetic data is that a change in a single dimension on a symbolic level usually corresponds to a multi-dimensional adaptation in acoustic, perceptual or articulatory space. To name a classic example, the abstract feature *voiced* may correspond to shorter or negative VOTs, actual presence of vocal cord vibrations, lower burst intensities, longer durations of a vowel preceding a voiced consonant etc. (e.g., Keating, 1984). Likewise, the impression of length – which is a phonologically relevant feature in many languages – may rely on acoustic aspects other than ob-

jective duration (Rosen, 1977; Cumming, 2011). Thus, typical phonetic investigations need to have a preconception about the phonetic parameters that need to be taken into account. In many domains such as the phonetic expression of emotion, voice pathologies, discourse phonetics or sociophonetics, we do have such clear-cut expectations, and chasing them may resemble the search for the metaphorical "needle in the haystack".

Following a similar proposition in (Malisz et al., 2019), we therefore suggest to make use of novel speech technological developments to assist solving (some of) the problems mentioned above. More specifically, we suggest the usage of recent methods in speech conversion and speech synthesis to support basic research in phonetics.

In the following, we give a first sketch of our envisaged methodological approach, and then discuss its potential usage as an exploratory and explanatory research tool within the general research programme of explainable artificial intelligence.

Methods

General idea

Our goal is to develop a neural generative model which purposely modifies one (out of many) specific dimension of phonetic variation, while leaving the remaining content of a speech signal unaltered. Ultimately, the methodology will thus enable us to manipulate the particular dimension we are interested in, e.g., dialect, speaker identity, gender, age, emotion, speaking style or phonetic or linguistic content, but keeping all those dimensions stable that may otherwise confound our analyses, e.g. verbal content. This *voice conversion* enables us to carry out a range of phonetic follow-up analyses: First, we are able to generate rich sounds we can use in perceptual studies. Typically, such stimuli are now created varying only very few acoustic-phonetic variables (e.g. pitch, formant patterns), and are often generated relying on more traditional speech synthesis algorithms that may introduce artefacts which may all by themselves influence the result of the study (Malisz et al., 2019).

Second, the generated signals lend themselves to in-depth acoustic-phonetic analyses, and help us to better understand and narrow down the extremely complex space of acoustic-phonetic parameters corresponding to the dimension of interest.

That way, our approach enables exploratory research, as the generated signals will help us narrow down the space of acoustic-phonetic parameters involved in the expression of phonetic variation (hypothesis generation). Lastly, they will also help us to explain phonetic variation by designing controlled follow-up experiments, extending or falsifying our existing theories and models.

Modeling Disentangled Dimensions of Speech Variation

In a first step towards this long-term goal (Gburrek et al., submitted), we developed an approach able to disentangle the dimensions of *speaker identity* and *linguistic content* in an utterance, thus being able to generate utterances containing the exact segmental structure (including fine phonetic detail) of one speaker with the voice characteristics of a second speaker. This approach goes beyond traditional approaches of voice conversion, as it does not need rely on recordings of identical linguistic content by several speakers.

This can be achieved by factorizing the speech representation into those traits to be converted and the remaining ones (e.g., Hsu et al., 2017). We follow a similar approach: a disentangled representation of the input speech signal is developed, where speaker characteristics are captured in one set of latent parameters, and content related variations in another. The starting point of our research is the factorized hierarchical variational

autoencoder (FHVAE, Hsu et al., 2017). In this approach, sources of variation are disentangled in a nonlinear low-dimensional latent space rather than in the observed data. A key assumption is that content induced properties of the speech signal, e.g. both the segmental and linguistically relevant suprasegmental structure, vary at a much faster rate than speaker-specific, para- and extra-linguistic factors. Rather, the latter are expected to remain relatively stable over time. This assumption is represented by a corresponding probabilistic graphical model in latent space, where a series of so-called *segment variables* capture short-term variations, supposedly caused by the linguistic content, and a series of *utterance variables*, which capture variations at a larger time-scale, supposedly caused by the speaker or environment characteristics present in the speech signal (Hsu et al., 2017).

The approach by Hsu and colleagues was extended by applying a convolutional neural network (CNN) based VAE encoder/decoder architecture, which allows to model short-term variations at a more fine-grained level of detail. The decoder reconstructs the segment variables with the utterance variables, combining the desired combination of speaker specific signal traits and linguistic content. The output of the FHVAE decoder are log-mel spectra, which are then synthesized using the WaveNet approach (Van den Oord et al., 2016). This has been shown to produce speech of extremely high quality, and often indistinguishable from human speech (Malisz et al., 2019). The WaveNet is trained independently of the target speaker. The technical details of the approach are described in Gburrek et al. (submitted).

A first voice conversion system is trained on TIMIT (Garofolo et al., 1993) and LibriSpeech (Panayotov et al., 2015) databases, and voice conversion is tested on new speakers, not seen during training.

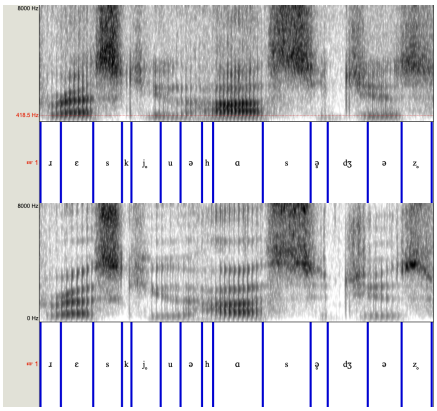

Figure 1: An example spectrogram of a reference utterance (top) and a target voice (bottom) producing the same linguistic content ("rescue hostages"). The spectral characteristics show a high degree of similarity in fine phonetic detail.

Results

We tested the applicability of the general approach for phonetic investigations with a detailed phonetic analysis. Specifically, we assessed the success in separating voice characteristics and linguistic content, or rather, segmental and suprasegmental linguistic content based on a comparison of three reference voices, each producing a different utterance, and two different target voices (1m, 1f).

A fine-grained narrow transcription of references and target voices was carried out, both based on an auditory impression and the acoustic signal. This analysis yielded a high degree of correspondence on the segmental level between reference and target voices. The target voices successfully mimicked the fine phonetic detail such as durational structure, formant trajectories, sound elisions and assimilations, and the fine-grained structure of plosives, including burst characteristics (cf. Fig. 1). Only in few cases, voice characteristics were not successfully reproduced, and in one instance, a sibilant [ʃ] sounded more [ç]-

like, possibly as a result of the target voice characteristics.

We believe that the treatment of pitch contours may be a crucial test case for our approach: Pitch contours both convey linguistic content by local, dynamically changing trajectories, but also carry plenty of information about speaker characteristics, such as global pitch level and range. Ideally, our approach should model the local trajectory changes of the reference signal, while simultaneously mimicking the global pitch characteristics of the target voice.

We compared the time normalized pitch contours of reference voices and voice targets with the conversion results. A comparison between pitch contours of reference voices, target voices and converged voices indeed shows that the pitch levels of the target voices are reproduced very successfully, while the local pitch trajectories more of less follow the dynamics of the reference contour. However, this comparison also shows occasional deviations between the converged and the reference contour (cf. Fig. 2). We therefore contend that in most cases, the conversion preserved the pitch contour with communicative relevance fairly successfully, but not perfectly, and was better at mimicking the global characteristics of the target voice.

An auditory impression yielded a very good imitation of the intended target voices, although the quality is still subject to some degradation introduced as part of the signal processing involved in the conversion procedure. In order to verify these impressionistic results in a more objective fashion, we calculated long term average spectra (LTAS) for the voiced parts of the reference signals, the target voices, and the converged voices. Here, the ideal case would be if the converged voice closely resembles the LTAS shape of the target voice. The results are encouraging (cf. Fig. 3), but occasionally inconclusive.

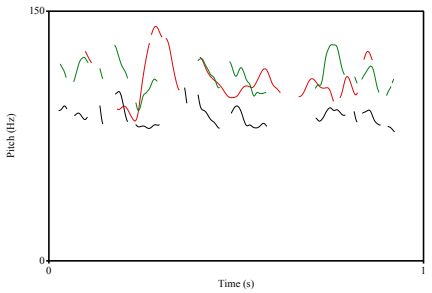

Figure 2: Comparison of a reference voice (black), the pitch contour of the target voice (red), and the pitch contour of the converged utterance. The green line has the global level and pitch range of the target voice, but mostly follows the pitch shape of the reference voice.

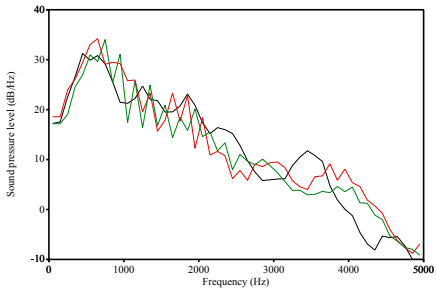

Figure 3: LTAS-based comparison of a reference voice (black), the target voice (red), and the converged voice (green). The converged voice shows a tendency to follow the pattern of the target voice rather than the reference voice.

A possible reason for this may lie in our methodological approach, as LTAS are not independent of the segmental content they are based on, and our material (a few sentences) may not have been sufficient for applying this method.

Discussion

Our investigations showed that it is indeed feasible to separate different dimensions of phonetic variation using state-of-the-art techniques of speech signal processing and synthesis. We plan to develop this approach further, to enable a separation of phonetic dimensions of interest that have hitherto been very dif-

ficult to grasp from a descriptive and explanatory perspective, due to their inherent underlying acoustic-phonetic complexity (e.g., speaker identity, age, gender, mood, dialect…) or their subtlety (e.g., early diagnosis of pathological voices). However, this project still needs work both in the technical realization of generating and synthesizing the phonetic variation underlying these dimensions, and in the development of methods to make the acoustic-phonetic aspects thus revealed truly interpretable. Such interpretability is needed in order to generate or enrich descriptive and explanatory models of phonetic variation, which could inform both experts (e.g., phoneticians, speech therapists) and laypersons who use their voices in professional contexts (e.g., teachers, actors).

We therefore see our ongoing project as being embedded in the overarching topic of explainable or interpretable artificial intelligence. More precisely, technological components are used as a research tool to provide cues for data exploration and (ultimately) explanation that help us extend or modify our existing theories and models of speech variation.

References

Cumming, R. 2011. The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics* 39(3), 375–387.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., & Dahlgren, N.L. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus (CDROM).

Gburrek, S., Glarner, T., Ebbers, J., Haeb-Umbach, R., & Wagner, P. (submitted). Unsupervised Learning of a Disentangled Speech Representation for Voice Conversion. *Proceedings of 10th Speech Synthesis Workshop (SSW10),* Vienna.

Hsu, W., Zhang, Y, & Glass, J. (2017). Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, arXiv:1709.07902.

Keating, P. (1984). Phonetic and Phonological Representations of Stop Consonant Voicing. *Language,* 60(2), 286-319.

Malisz, Z., Henter, G.E., Valentini-Botinhao, C., Watts, O., Beskow, J., & Gustafson, J. (2019). Modern Speech Synthesis for Phonetic Sciences: A Discussion and an Evaluation. *Proceedings of ICPhS 2019*, Melbourne, Australia.

Panayotov, V. Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Pierrehumbert, J. (2002). *Word-specific phonetics*. In Carlos Gussenhoven and Natasha Warner (eds.), *Laboratory Phonology* 7, 101-139. Berlin & New York: Mouton de Gruyter.

Rosen, S. M. 1977. The effect of fundamental frequency patterns on perceived duration. In: *Speech Transmission Laboratory—Quarterly Progress and Status Report* volume 18. Stockholm, Sweden: KTH 17–30.

Roy, N., Barkmeier-Kraemer, J., Eady, T., Preeti Sivasankar, M., Mehda, D., Paul, D., & Hillman, R. (2013). Evidence-based systematic clinical voice assessment: a systematic review. *American Journal of Speech-Language Pathology,* 00, 1-18.

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N. Senior, A. and Kavukcuoglu, C. (2016). WaveNet: A generative Model for Raw Audio, arXiv:1609.03499.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1-12. doi:10.1016/j.wocn.2014.11.001