Do you remain the same speaker over 21 recordings?

Nicolas Audibert, Cécile Fougeron, Estelle Chardenon

Do you remain the same speaker over 21 recordings?

Nicolas Audibert, Cécile Fougeron and Estelle Chardenon
Laboratoire de Phonétique et Phonologie
UMR7018 CNRS/Sorbonne Nouvelle, Paris, France
{nicolas.audibert, cecile.fougeron, estelle.chardenon}@sorbonne-nouvelle.fr

Introduction: In forensics, an open question concerns the validity of a comparison between recordings weeks, months, or years apart, and which conditions allow such comparisons. While variation between speakers or speech conditions has been the focus of many phonetic studies, our knowledge on intra-speaker variability across multiple recordings of the same task is surprisingly very limited. Among other factors, age (Biever and Bless, 1989; Jacewicz, 2009), quality of life (Campbell, 2009; Verdonck, 2004), fatigue and emotional state (Scherer et al, 1998; Hollien, 1990), as well as the speech task (Dellwo, 2015) or communicative situation (Scarborough and Zellou, 2013) are known to induce differences in the speech produced by the same individual. More recently, Chardenon (2020) showed that intra-speaker variability on temporal dimensions is larger between distant recording than between successive recordings. This work is part of a larger project on methodological issues in voice comparison, with a specific focus on intra-speaker speech variation across multiple recordings and on the effect of the time lapse between recordings. In the study presented here, we evaluate on a limited set of speakers recorded multiple times over 7 years, whether we can observe the emergence of speaker specific profiles of variation when looking at selected speech dimensions.

Method: Ten speakers were recorded each year three times in a row over a period of seven years, and twice a week over a 1-month period on the same speech material enabling controlled phonetic comparisons between the 29 recording sessions. Recorded material includes read and spontaneous speech as well as other speech-like tasks, in a protocol meant to investigate multiple dimensions of speech and voice. The results presented here are based on 5 female and 3 male speakers, all French native, aged 39 to 58 years old at the date of the first recording, living in the region of Paris and belonging to the same social and professional category. We used 18 to 21 of their recordings (3 successive recordings each year during 6 or 7 years) on the reading of the French version of the tale 'The North wind and the sun'. The text was divided in18 predefined chunks of 15 to 24 phonemes each. All 159 recordings were manually segmented into these 18 chunks, as well as in pauses (with a threshold of 200 milliseconds) and speech. Six features were extracted on each chunk using a Praat script. Information related to the temporal organization of speech is captured over 3 domain-sizes through measures of (i) speech rate (with pauses), (ii) articulation rate (without pause), and (iii) a 'voiced ratio' defined as the total duration of voiced segments over the speech duration. Mean speaking F0 and F0 range over each chunk (in semitones) capture information related to voice and intonation, and the slope of the LTAS captures spectral information related to both laryngeal and supra-laryngeal activities. For these six features, in addition to mean values computed per chunk, the fluctuation of a speech feature is estimated by computing normalized differences (hereafter d(featureX)) between consecutive chunks as $|chunk_i - chunk_{i-1}|/((chunk_i + chunk_{i-1})/2)$. Each recording is thus characterized by 12 features.

Results and discussion: In order to test for the effect of the speaker identity, the recording session, and their interaction, a linear mixed effects model was fitted using the lme4 package (Bates et al., 2015) for each feature converted to z-scores, with the chunk identifier as random intercept. While all 12 features except d(LTAS slope) are found to be speaker dependent, a significant effect of the recording session is found only on mean values per chunk. While differences between consecutive chunks vary by speaker and by recording independently, a recording by speaker interaction is found for mean values by chunk. Interestingly, the descriptors that are more stable across recordings are the ones linked to the fluctuation of the speech dimensions from one chunk to the next over the recording.

In order to further understand individual profiles of variation between recordings, normalized variation levels were estimated by computing the standard deviation over all recordings for each speaker on the 12 features, after conversion to z-scores. As illustrated on Figure 1, variation in temporal dimensions (speech and articulation rate as well as voiced ratio) appears as rather homogenous across speakers with similar variability levels for all of them except articulation rate of M01 and M03. Discrepancies between patterns observed for speech and articulation rate can be attributed to differences in the variability of

pauses number and duration, used by some speakers to compensate variations in articulation rate. On the other hand, speakers are far more different from each other regarding features linked to local variation of F0 (variation of F0 range and fluctuation of F0 mean and range). Although larger differences are found on F0-related features between female speakers, no clear sex-specific patterns are observed.

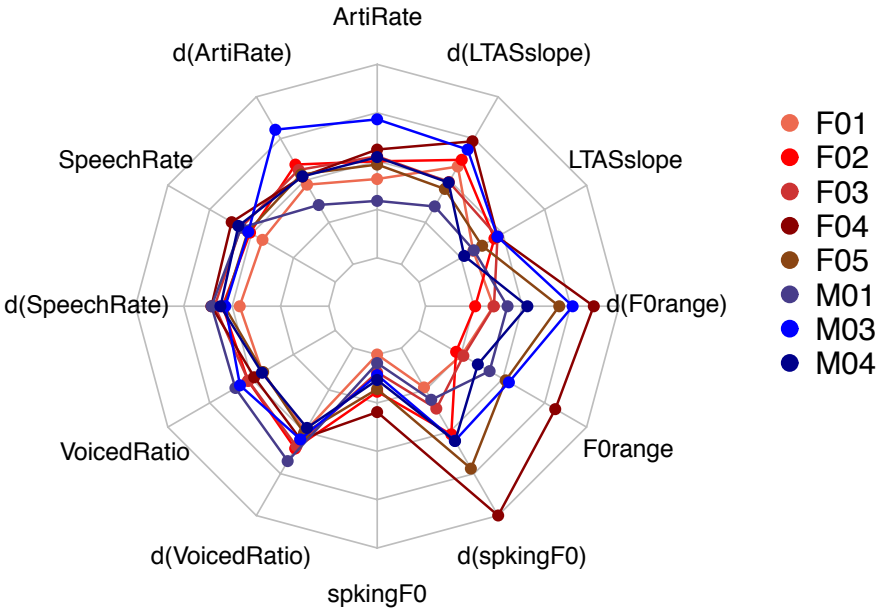

Figure 1: Variation profiles between recording sessions of the 8 speakers, on the 12 speech features expressed as normalized standard deviation for the 6 features measured on the 18 chunks per recording, and their fluctuation between consecutive chunks (chunk to chunk differences are noted as 'd(feature)').

These preliminary results on 8 speakers on the same reading task suggest that measures of fluctuation of speech timing and F0 between consecutive chunks depend more on the speaker than on the repetition. Such measures of local variability, taken at a larger time-scale than the first and second derivatives classically used in automatic classification tasks, may be useful to improve the robustness of speaker identification. They also suggest that patterns of variability between recording sessions are speaker-dependent, particularly on F0 related features. Data recorded for the remaining 2 speakers (1 male, 1 female) are currently being analyzed while the extension of such analysis to spontaneous productions of the same speakers is being investigated. Further analysis will be carried out with more features and a more comprehensive description of the prosodic phrasing of the text in each recording.

References

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).

Biever, D. M., Bless, D. M. (1989). Vibratory characteristics of the vocal folds in young adult and geriatric women. *Journal of Voice*, 3(2), 120-131.

Campbell, W., Bonastre, J-F., Schwartz, R., Driss, M. (2009). Forensic Speaker Recognition. *Signal Processing Magazine*. 26.2: pp-95-103.

Chardenon, E., Fougeron, C., Audibert, N., Gendrot, C. (2020). Dis-moi comment tu varies ton débit, je te dirai qui tu es. *31e Journées d'Études sur la Parole*. Nancy, France, 82-90.

Dellwo, V., Leemann, A., Kolly, M-J. (2015). Rythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *JASA*, 137(3): 1513-1528.

Harrington, J., Palethorpe, S., & Watson, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. *Interspeech 2007*.

Hollien, H., 1990. The acoustics of Crime. (1990). *The New Science of Forensic Phonetics*. Dordrecht: Springer.

Jacewicz, E., Fox, R-A., O'Neill, C., Salmons, J. (2009). Articulation rate across dialect, age and gender. *Lang Var Change*. 1; 21(2): 233–256.

Scarborough, R., & Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *JASA*, 134(5), 3793-3807.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). *Vocal expression of emotion*. Oxford Univ. Press.

Verdonck-de Leeuw, I. M., & Mahieu, H. F. (2004). Vocal aging and the impact on daily life: a longitudinal study. *Journal of Voice*, 18(2), 193–202.