Individual specificity, redundancy and the evolution of phonological systems:
Intonation in a tone language
Francesco Cangemi¹, Christian Weitz², Kiều Phương Hạnh¹, Marc Brunelle³, Martine Grice¹
¹ILF Phonetik, University of Cologne; ²University of Ottawa; fcangemi@uni-koeln.de

Background. Speakers encode phonological contrasts redundantly, i.e. using multiple cues (e.g. Lisker 1986) which are distributed across different time domains (e.g. Coleman 2003). Such redundancy has been recently seen as a key feature in both ontogenetic and phylogenetic mechanisms of language evolution, i.e. both over the timescale of decades in individual acquisition (Shattuck-Hufnagel et al. 2015) and over the timescale of centuries in language change (Winter 2014). Moreover, when multiple cues are available to speakers, different speakers are known to capitalize on different cues (e.g. Repp 1983, Niebuhr et al. 2011). This suggests that individual-specific strategies in the encoding of contrasts might be more frequent when contrasts are particularly redundant (RESH). Stretching this line of thoughts to its extreme, individual specificity in phonetic encoding might be expected to be particularly pervasive for contrasts which are usually encoded through non-phonetic devices, but which also have incipient, fading or non-grammaticalised phonetic encoding. For this reason, as an initial testing ground for RESH, we take the case of phonetic encoding of non-lexical or sentence-level contrasts in a tone language, where they are usually encoded through the use of non-phonetic (e.g. morphosyntactic) devices.

Northern Vietnamese has a tone system with six lexical tones which is largely based on f0 and voice quality (e.g. Thompson 1965, Pham 2003 and Michaud 2004). Non-lexical meaning is instead largely conveyed through sentence final particles (e.g. Thompson, 1965), as in the case of di for imperatives. In the last twenty years, however, a number of studies have documented phonetic encoding of non-lexical meaning, through the interplay of pitch, duration, intensity and voice quality (e.g. Đỗ et al. 1998 for sentence modality, Michaud and Vũ 2004 for emphasis, Jannedy 2007 for information structure), thus documenting intonational phenomena in Northern Vietnamese. Recent research on spontaneous interaction shows that some discourse functions (e.g. backchannel, repair initiation) are not only encoded intonationally, but that the intonational use of f0 in these contexts can even override the expected tonal patterns (Hạnh & Grice 2010). Initial observations of written discourse have shown that internet users even apply tonal diacritics to express discourse-level meaning. Brunelle et al. (2012) show with read speech data that speakers encode sentence modality (question vs. statement) and affect (neutral vs. emphatic) contrasts by using continuous phonetic cues (overall intensity, pitch or duration), discrete phonological options (intonational tones), or a combination of both, but also report highly speaker-specific behaviour. Similarly, Brunelle (2015) suggests for Southern Vietnamese that while sentence modalities have marginally different intonational properties, it is not possible to argue for a fully grammaticalised intonational system.

Rationale. Given this background on the interplay of tone and intonation in Northern Vietnamese, RESH would predict that if in this language non-lexical contrasts are encoded intonationally at all, a high degree of speaker-specificity is to be expected. More specifically, the pervasiveness of speaker-specific effects in intonational encoding of non-lexical contrasts is predicted to be higher in Northern Vietnamese than in languages which do not encode such contrasts as redundantly (i.e. in languages without sentence final particles). However, the very possibility of testing these predictions relies on the availability of in-depth methods for profiling of speaker-specific strategies. Such methods are however not widespread (but see Drager & Hay 2012, Cangemi et al. 2015). In this submission we develop an infrastructure for the quantification of speaker-specific variability, and use it to provide an in-depth analysis of the Brunelle et al. (2012) Northern Vietnamese corpus.

Method. The dataset features 72 sentences ending with không (lexically high-level, meaning ‘empty/only’ or ‘yes/no question particle’), recorded by 16 speakers of Standard Vietnamese in four different communicative functions (statement, question, annoyed statement and insisting request) yielded by the combination of sentence modality (question vs. statement) and affect (neutral vs. emphatic). The sentences consist of four syllables and are tonally as well as segmentally identical. We first analyse the dataset by profiling each speaker with respect to whether duration and/or f0 excursion of the target word không are reliably used to encode the modality and/or affect contrasts (top-down analysis, see Fig. 1). The results are then used to predict the outcome of an unsupervised clustering algorithm, ran on data from the four meaning categories (bottom-up analysis). The solutions for each
individual speaker (in terms of the optimal number of clusters and in the categorization errors made by the algorithm, see Fig. 2) are then interpreted with respect to the profiling from the top-down analysis.

**Results.** Fig. 1 shows for each speaker (rows) and for each cue (panels, further subdivided in nested contrasts: columns) the normalised absolute distance between categories (greyscale, with darker shades indicating larger values, full black indicating 2 standard deviations). For example, the dark cell indicated by the red arrow shows that speaker F1 productions for emphatic statements and questions are largely different in terms of f0 excursion, whereas this is definitely not the case for speaker F2. This representation of the top-down results already provides an intuitive display of speaker-specific strategies, detailing which speakers use which cues to encode which contrasts. Crucially, it is complemented by plots for individual speakers from the bottom-up analysis, as illustrated in Fig. 2. Here the individual data points from speaker F3 are symbol-coded for design categories (questions: triangles, statements: circles; neutral: empty, emphatic: filled), colour-coded for unsupervised clustering categorization, and plotted on the (normalised duration, normalised f0 excursion) plane. *Prima facie*, no reliable encoding of modality or affect is suggested for speaker F3, neither via the top-down analysis (white cells) nor via the clustering solution (mixture of different symbols for a same colour). However, by interpreting the two displays jointly, it is possible to see that data from speaker F3 is actually quite structured along the sentence modality dimension, with statements having relatively consistent slight f0 falls (circles in the centre of the plot, around y=−1) but with questions having either slight f0 rises or sharp f0 falls (triangles in both the top and bottom plot region), and with a somewhat reversed picture for duration, with questions having relatively stable duration (triangles in the centre of the plot, around x=0) and statements having a larger duration range (circles spanning the plot from left to right).

This crossed pattern is only attested in the entire corpus for speaker F3. As can be inferred from Fig. 1, other speakers make use of different but equally interesting specific strategies in the encoding of sentence modality, affect or both. However, although they are unusual and difficult to detect in the absence of appropriate analysis tools, these patterns document the presence of intonation in a tone language, and suggest that speaker-specific effects might indeed be pervasive in the phonetic encoding of contrasts usually expressed through other devices such as final particles, making the results compatible with RESH.