INFLUENCES OF SPEAKER ATTITUDES ON GLOTTALIZED TONES: A STUDY OF TWO VIETNAMESE SENTENCE-FINAL PARTICLES

Dang-Khoa Mac[1], Thi-Lan Nguyen[1], Alexis Michaud[1, 2], Do-Dat Tran[1]

[1] International Research Institute MICA, HUST – CNRS/UMI-2954 – Grenoble INP,
Hanoi University of Science and Technology
[2] Langues et Civilisations à Tradition Orale (LACITO), CNRS – Univ. Paris 3 Sorbonne Nouvelle
{ Dang-Khoa.Mac, Thi-Lan.Nguyen, Alexis.Michaud, Do-Dat.Tran }@mica.edu.vn

ABSTRACT

Attitudinal information in a spoken utterance can be lexically encoded; it can also be conveyed by intonation, including modification of voice quality. This study aims to investigate how speaker attitude affects the realization of glottalized tones in Vietnamese. A specific recording setup was designed; a corpus containing attitudinal sentence-final particles (SFPs) was recorded by ten speakers. The present report contains qualitative observations and quantitative assessments for two glottalized tones (B2, the “drop tone”, and C2, the “broken tone”) in utterances realized with attitudes of SURPRISE or IRRITATION, as contrasted with simple DECLARATION. The results suggest that there is a considerable range of intonational (allotonic) variation in the realization of glottalization, and that it contributes to expressing speaker attitude – and is not unlikely to convey other types of prosodic information as well.

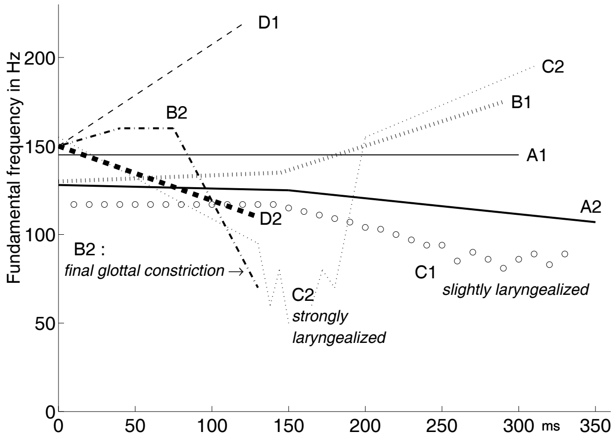
Keywords: Vietnamese, glottalization, tones, attitudes, sentence-final particles, expressive speech

1. INTRODUCTION

Speech conveys information about mental, intentional, attitudinal and emotional states. In recent years, special attention has been paid to voice quality (phonation types) in the field of expressive speech research [2, 3, 26]; voice quality has been described as a fourth dimension of prosody, in addition to the classical parameters of F_0, intensity, and duration ([7]; see also [4, 12, 21]).

In Northern Vietnamese, a tonal language, F_0, intensity, duration and phonation types also play a role at the phonemic level, i.e. for lexical access (for perceptual confirmation: [5, 14]). Vietnamese has 6 tones on smooth syllables (syllables without a final stop), represented in Figure 1: level (A1), falling (A2), rising (B1), drop (B2), curve (C1), and broken (C2). This system involves contrastive use of phonation types. Specifically, glottalization occurs during the production of C2 (called *ngã* in Vietnamese) and B2 (*nặng*). C2 is a rising tone with strong glottalization in its first half. In typical realizations of tone B2, F_0 is initially in the speaker’s mid-range; it then falls dramatically because of a strong glottalization in its second half. It appears worthwhile to investigate how speaker attitudes affects the realization of Vietnamese glottalized tones: how “linguistic” and “paralinguistic” dimensions combine and interact to shape observed patterns of glottalization.

Figure 1: Schematic diagram of Hanoi Vietnamese tones (from [19], with permission)



A pilot study [22] reported observations about Vietnamese sentence-final particles (hereafter SFPs), which carry both lexical tone and attitudinal information. Significant differences in F_0, duration, and voice quality were brought out by comparing the realizations of the same SFP in sentences uttered with an attitude of SURPRISE or IRRITATION, as compared with simple DECLARATION.

The extension of the work reported in the present paper aims to quantify measurements over a larger dataset than was used in the pilot study.

2. DESIGN OF THE CORPUS

2.1. Speech materials: selection of particles and attitudes, and specification of contexts

As in the pilot study [22], the key materials in the corpus consist of sentences with SFPs carrying different lexical tones. The choice of speaker attitudes was made among the set of sixteen distinguished in a study of Vietnamese attitudes [16]. On the basis of that study's perceptual results, seven attitudes were selected because they had been found to be best distinguished by listeners on the basis of the audio (i.e. without visual information) [17]. They are: DECLARATION (DEC), INTERROGATION (INT), SURPRISE (SUR), OBVIOUSNESS (OBV), IRRITATION (IRR), AUTHORITY (AUT), and SARCASTIC IRONY (SAR). The carrier sentences used here are shown in Table 1.

Table 1: Target sentences in the corpus. SFPS are in **bold**. Indications on the attitude conveyed by the SFP are added in square brackets.

Sentences	Meaning	SFP tone
Ba đi học **ạ**.	Ba goes to class. [Politeness]	drop (B2)
Ba đi học **đã**.	First, Ba goes to class.	broken (C2)
Ba đi học **mà**.	Ba goes to class, of course. [OBVIOUSNESS]	falling (A2)
Ba đi học **hả**?	So Ba is going to class? [SURPRISE]	curve (C1)

Further limitations were made to the scope of the study in view of restrictions on combinations among attitudes, among SFPs, and between attitudes and SFPs. The semantics of a SFP can make it quasi-indispensible to the expression of a certain attitude, and conversely, render it largely incompatible with other attitudes. For example, the SFP *ạ* conveys politeness and deference, so that highly specific contexts are necessary for its occurrence with an attitude of sarcastic irony or authority. As a consequence, the data set is not fully symmetrical. No attempt was made here to combine SFPs (i.e. to have two or more SFPs in the same sentence, as can happen in the language). Four SFPs and seven attitudes were chosen; seventeen suitable combinations among those were selected, and fleshed out as "target sentences" embedded in contextualized dialogues.

2.2. Speakers and corpus recording

Twenty (paid) Vietnamese speakers (10 male and 10 female; mean age: 21) participated in the recordings; only data from the 10 male speakers are reported here. Beyond the minimal requirement of selecting native speakers of Northern Vietnamese, we recruited them from the Hanoi Academy of Theatre and Cinema *(Đại học Sân khấu Điện ảnh)*, with a view to obtaining an appropriately expressive rendering of the materials. The results of the pilot study suggested that students in engineering (the default pool of speakers for recordings at our University) find it challenging to perform different attitudes with the desired degree of clarity and consistency.

Detailed contextualization of each target sentence was provided (in writing), and the target sentences were embedded in a dialogue, in order to guide the participants' interpretation of the task with as much precision as possible [23]. One of the investigators sat in the recording booth and served as interlocutor; the investigator and consultant were visually separated by a curtain. Consultants were allowed to repeat each target sentence (always with a dialogue partner) until they were satisfied with their performance, at which point they repeated the target sentence at least two times. They went through the data set twice; in total, the recording lasted about 30 minutes for each speaker. The amount of recorded data was large; among repetitions of a given target sentence, the six "best" performances (for each speaker) were chosen by one of the investigators (a native speaker with experience of research into expressive speech). In total, the corpus for analysis contained 2040 utterances (20 speakers * 17 target sentence * 6 samples); only part of the data are analyzed here. (Data analysis through formal perceptual methods has not been conducted so far; this topic is touched upon in the Conclusion.)

The speaker's voice was recorded in stereo, with (i) a high-quality head-worn microphone and (ii) a microphone placed about 50 cm from the mouth of the speaker. An electroglottographic (EGG) signal was simultaneously collected, in order to obtain indications on vocal fold contact area [9, 11]; the equipment used was a two-channel EGG [24]. The files comprise four channels (WAV, 44,100 Hz, 24-bit): two audio channels from the speaker; one audio channel from the interviewer; and the EGG signal. The files were segmented and annotated at the level of the sentence and the syllable.

3. DATA ANALYSIS

The corpus was designed for the needs of research into Vietnamese expressive speech; the present paper only exploits part of the data, to study the influence of attitudes on glottalized tones. The topics of research are the following:

- realizations of the drop tone (tone B2) in association with SURPRISE
- realizations of the broken tone (tone C2) in association with IRRITATION.

The influence of these attitudes is brought out through comparison with realizations of segmentally identical sentences realized with an attitude of DECLARATION – an attitude which for our purposes serves as a point of reference [1], although it clearly allows for various nuances.

3.1. Analysis method for glottal phenomena

The method used is based on the derivative of the EGG signal [13]. F_0 and the glottal open quotient O_q were estimated with PEAKDET, a script available from the COVAREP GitHub repository [8].

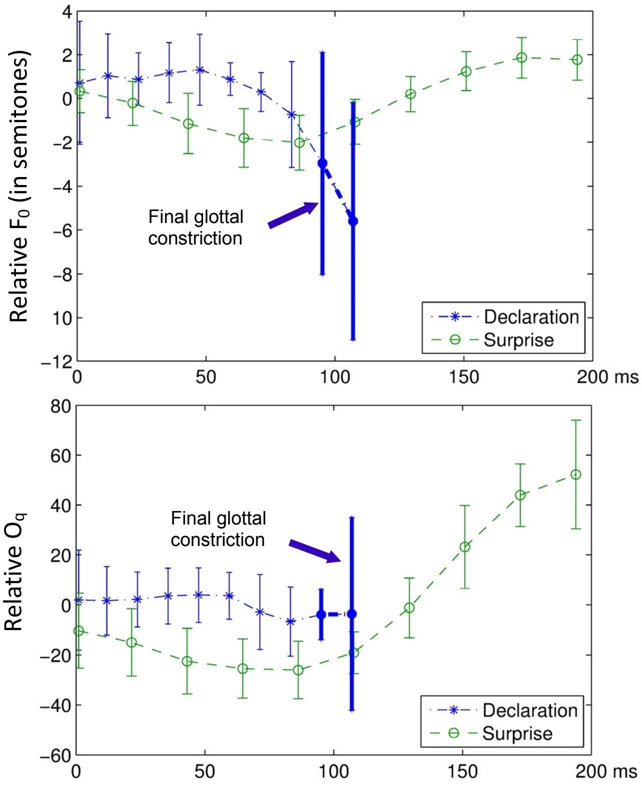
For automatic detection of creaky voice, i.e. vibratory mechanism zero [25], a Matlab script was developed. If variation in cycle length from one glottal cycle to the next (ΔF_0) is alternatively positive and negative for adjacent cycles, and is higher than 15%, the cycles at issue are labelled as "creaky". Note that "pressed voice" refers to modal voice with relatively stronger vocal fold adduction than "breathy/whispery voice"; from the point of view of vocal fold physiology, "pressed", "modal" and "breathy/whispery" voice are found along a continuum [25], and hence no attempt was made to categorize tokens according to essentially arbitrary thresholds of open quotient.

F_0 and O_q were recalculated with reference with mean values for each speaker, yielding relative F_0 values in semitones, and relative O_q values in %, that can be averaged (or compared) across speakers. The formulas used are those proposed in [18].

3.2. SURPRISE and the drop tone (tone B2)

Figure 2 presents average F_0 and O_q contours for the SFP *ạ* [IPA: /a/], which bears the drop tone (tone B2), in segmentally identical utterances that contrast in terms of attitude (SURPRISE vs. DECLARATION); the data are from the 10 male speakers (60 tokens).

Figure 2: Average F_0 and O_q contours for tone B2



Compared to DECLARATION, tone B2 with SURPRISE has twice longer duration. As for the F_0 contour, B2 in DECLARATION has the telltale characteristics reported in the literature as its standard template [6, 15]: it is rather flat during its initial half, then decreases to a final sharp drop. With SURPRISE, the F_0 contour is strikingly different. An early dip, extending over one third of its duration, is followed by an increase to a relatively high value, with the maximum located at the end (which, due to the final position of the SPF in the utterance, corresponds to the offset of voicing).

In terms of glottalization, differences are also salient. The glottalization that is part of the tone's specification is present under both conditions, confirming its robustness as a cue to tone [20]. In terms of phasing, however, glottalization shifts to early/medial position in SURPRISE (with a dip in O_q, down to values characteristic of *pressed voice*) instead of its canonical alignment with the end of the syllable (as observed in DECLARATION, where the high standard deviation of F_0 and O_q at the end constitutes strong evidence of glottal constriction).

3.3. IRRITATION and the broken tone (tone C2)

Figure 3: Average F_0 and O_q contours of 10 male speakers (60 tokens) on tone /C2/

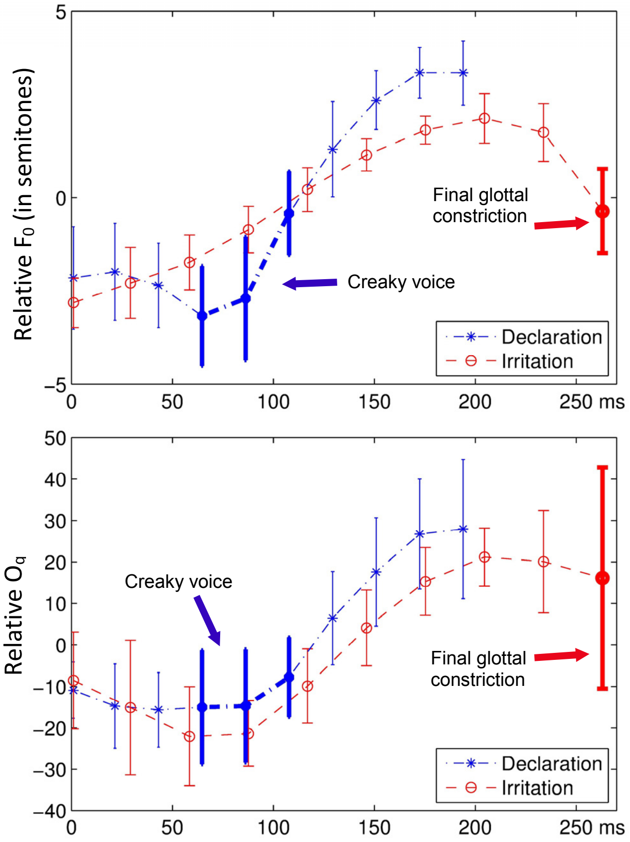


Figure 3 illustrates differences in the realization of tone C2 in association with DECLARATION and IRRITATION from 10 male speakers. Measurements are conducted over the vowel /a/, exclusive of the initial (preglottalized /ɗ/), as tone is carried by the syllable rhyme [5, 28].

The shape of F_0 contours differs between the two attitudes. The dipping in the middle of F_0 contour found in DECLARATION (a telltale characteristic of this tone as described in the literature [27, 30]) is absent in IRRITATION, where F_0 increases gradually until about 75% of the vowel's duration. Duration is slightly greater in IRRITATION than in DECLARATION.

As for O_q, tone C2 in DECLARATION starts with modal voice, lapses briefly into glottalized voicing, and ends with modal voice. In Figure 3, the glottalized portion is set out visually by means of thicker lines; this glottalization is clearly medial, as mentioned in converging reports about this tone's canonical realization [27, 30]. In IRRITATION, tone C2 shows evidence of glottal constriction in *two* positions: medial (as evidenced by the dip in O_q in Figure 3), and final. The former is more consistently present, and can be interpreted straightforwardly as an intonational variant of the tone's medial glottalization. As for the latter, it bears some phonetic resemblance to the final glottal constriction found in tone B2, but its functional origin is altogether different: it is not part of the lexical tone. It is part of intonation (in the sense of [29]): a 'hard' offset of voicing (final glottal closure) is cross-linguistically associated to 'harsh/brutal' attitudes [10].

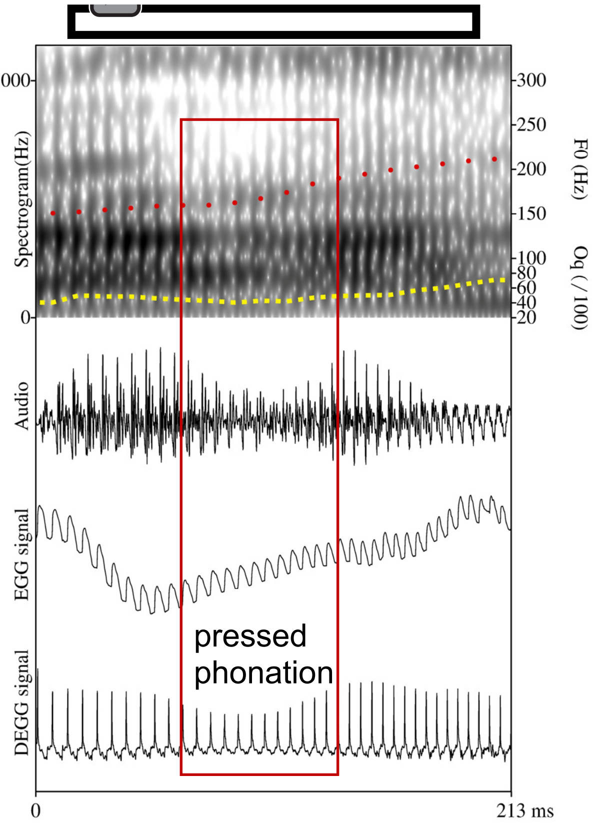
4. CONCLUDING NOTES AND PERSPECTIVES

The qualitative analysis results reveal the considerable effects of attitude on the realization of Vietnamese glotalized tones. Highlights already reported in [19], and confirmed here, are (i) the realization that tone B2's final glottal constriction can be realized early on the syllable (and as pressed voice, rather than a lapse into creaky phonation proper) in SURPRISE, and (ii) the *double* glottalization of tone C2 in IRRITATION, due to an added ("attitudinal") final glottal constriction.

Additionally, when sifting through the data, unexpected examples were found. A case in point is the realization of tone C2 with pressed voice (instead of the expected lapse into creaky voice) in DECLARATION, where one would expect realization of the tone to be closest to the canonical template. An example is shown in Figure 4.

Such examples call for closer scrutiny of the attitude of DECLARATION; we hypothesize that the strength of glottalization for a C2-tone SFP correlates positively with the degree of expressed finality.

Figure 4: Example (with embedded sound) of pressed voice in tone C2, attitude DECLARATION (speaker M11)



The mid-term goal consists in providing a statically analysis to investigate the common characteristics of the Vietnamese glottalized tones. The present results offer some insights into the considerable phonetic range covered by Vietnamese glottalized tones. Overall, they confirm earlier findings [22], on a stronger empirical basis (60 tokens for each tone*attitude combination). These results obtained on production data now call for (i) statistical data treatment, covering all the materials recorded, and (ii) perceptual verification: labels such as 'IRRITATION' have so far been used as a matter of definition instead of obtaining a characterization through perceptual procedures. The long-term agenda consists in arriving at a multi-parametric model of the interplay between speaker attitudes and the complex lexical tones of Vietnamese.

5. ACKNOWLEDGMENTS

The authors are grateful to Ji Xi for his proofreading. This work is supported by the Vietnamese government project KC.03.07/11-15 (Researching, Designing and Building intelligent automatic systems using human-machine interaction by Vietnamese speech in noise environment); it also represents a contribution to the "Phonetic and phonological complexity" component of LabEx "Empirical Foundations of Linguistics" (ANR-10-LABX-0083).

6. REFERENCES

[1] Aubergé V. 2002. A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. Proc. Speech Prosody 2002. Aix en Provence, pp 151–155.

[2] Audibert N., Aubergé V., Rilliard A. 2005. The relative weights of the different prosodic dimensions in expressive speech: A resynthesis study. Proc. ACII 2005. Springer, Beijing, pp 527–534.

[3] Banse R., Scherer KR. 1996. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70:614–636.

[4] Bänziger T., Scherer KR. 2005. The role of intonation in emotional expressions. Speech Commun 46:252–267.

[5] Brunelle M. 2009. Tone perception in Northern and Southern Vietnamese. J Phon 37:79–96.

[6] Brunelle M., Nguyễn Khắc Hùng, Nguyễn Duy Dương. 2010. A Laryngographic and Laryngoscopic Study of Northern Vietnamese Tones. Phonetica 67:147–169.

[7] Campbell N., Mokhtari P. 2003. Voice quality: the 4th prosodic dimension. Proc. ICPhS XV. University of Barcelona, Barcelona, pp 2417–2420.

[8] Degottex G. COVAREP: A Cooperative Voice Analysis Repository for Speech Technologies. https://github.com/covarep/covarep. Accessed 12 Feb 2014.

[9] Fabre P. 1957. Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence. Bull Académie Natl Médecine 141:66–69.

[10] Fónagy I. 1983. La vive voix: essais de psycho-phonétique. Payot, Paris.

[11] Fourcin A. 1971. First applications of a new laryngograph. Med Biol Illus 21:172–182.

[12] Gobl C., Ní Chasaide A. 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Commun 40:189–212.

[13] Henrich N., d' Alessandro C., Castellengo M., Doval B. 2004. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. J Acoust Soc Am 115:1321–1332.

[14] Kirby J. 2010. Dialect experience in Vietnamese tone perception. J Acoust Soc Am 127:3749–3757.

[15] Kirby J. 2011. Vietnamese (Hanoi Vietnamese). J Int Phon Assoc 41:381–392.

[16] Mac D.-K. 2012. Génération de parole expressive dans le cas des langues à tons. Ph.D., Université de Grenoble.

[17] Mac D.-K., Aubergé V., Rilliard A., Castelli E. 2010. Cross-cultural perception of Vietnamese audio-visual prosodic attitudes. Proc. Speech Prosody 2010.

[18] Mazaudon M., Michaud A. 2008. Tonal contrasts and initial consonants: a case study of Tamang, a "missing link" in tonogenesis. Phonetica 65:231–256.

[19] Michaud A. 2004. Final consonants and glottalization: new perspectives from Hanoi Vietnamese. Phonetica 61:119–146.

[20] Michaud A., Vu-Ngoc T. 2004. Glottalized and nonglottalized tones under emphasis: open quotient curves remain stable, F0 curve is modified. In: Bel B., Marlien I. (eds) Speech Prosody 2004. Nara, Japan, pp 745–748.

[21] Mozziconacci S. 2002. Prosody and emotions. Proc. Speech Prosody 2002.

[22] Nguyen T.-L., Michaud A., Tran D.-D., Mac D.-K. 2013. The interplay of intonation and complex lexical tones: how speaker attitudes affect the realization of glottalization on Vietnamese sentence-final particles, Proc. Interspeech 2013, Lyon, France.

[23] Niebuhr O., Michaud A. 2015. Speech data acquisition: the underestimated challenge. KALIPHO - Kieler Arbeiten Zur Linguist und Phonetik 3:1–42.

[24] Rothenberg M. 1992. A multichannel electroglottograph. J Voice 6:36–43.

[25] Roubeau B., Henrich N., Castellengo M. 2009. Laryngeal vibratory mechanisms: the notion of vocal register revisited. J Voice 23:425–38.

[26] Shochi T., Aubergé V., Rilliard A. 2007. Cross-listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes. Proc. ICPhS XVI. pp 2097–2100.

[27] Thompson L.C. 1984. A Vietnamese Reference Grammar. Mon-Khmer Stud. 13-14.

[28] Tran D.-D. 2007. Synthèse de la parole à partir du texte en langue vietnamienne. Thèse en cotutelle international MICA.

[29] Vaissière J. 2004. The Perception of Intonation. In: Pisoni D.B., Remez R.E. (eds) Handb. Speech Percept. Blackwell, Oxford, U.K. & Cambridge, Massachusetts, pp 236–263.

[30] Vu-Ngoc Tuân, d'Alessandro C., Michaud A. 2005. Using open quotient for the characterization of Vietnamese glottalized tones. Eurospeech-Interspeech 2005 9th Eur. Conf. Speech Commun. Technol. Lisboa, pp 2885–2889.