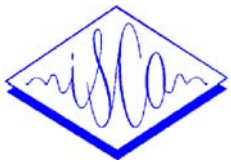

International Symposium on Tonal Aspects of Languages:
With Emphasis on Tone Languages
Beijing, China, March 28-31, 2004

Pitch and Alignment in the Perception of Tone and Intonation: Pragmatic Signals and Biological Codes

David House

Department of Speech, Music and Hearing, Centre for Speech Technology
KTH, Stockholm, Sweden
davidh@speech.kth.se

Abstract

In this paper a case is made for examining tone perception and particularly phenomena of tonal alignment in terms of perceptual limitations to pitch processing. The perception of pitch and tonal movement is discussed in relationship to research on alignment. Results from perception experiments using Swedish listeners and non-scripted speech production data from Swedish speakers are used to exemplify the discussion. The results of these experiments are further discussed in terms of biological codes. Finally thresholds of tonal movement perception are proposed based on principles of syllable alignment consistent with perceptual constraints including cognitive processing.

1. Perceptual constraints in tonal processing

There is currently considerable interest in examining the relationship between tone and intonation in diverse languages of the world. With its lexical tone inventory and long history as a language studied by scholars, Standard Chinese is a language well suited for investigating tonal phenomenon and the interactions between tone and intonation. Professor Zongji Wu, whom we honour with this symposium, has carried out much pioneering and inspirational work on this challenging topic [1]. In his analysis, he sees the surface realization of intonation in Standard Chinese as a result of the interaction of three components. Two of the components make up basic tone units. These are the original mono-syllabic toneme patterns and the poly-syllabic phrasal tone sandhi. The third component is sentential intonation reflecting declarative, interrogative or exclamatory intonation. The intonation component modifies the global contour and can also be changed by different speaker attitudes. He sees intonation as changing the register of the global tonal contour in much the same way as a change of key in music.

The analysis of intonation as a result of the interaction of underlying components as formulated by Professor Zongji Wu helps us understand the tonal variation found in production studies for both scripted and non-scripted databases [2]. This kind of analysis is also useful in the generation of the F0 contour in speech synthesis, especially as we try to expand speech synthesis to include expressive speech under a variety of emotional situations. The complexity of intonation especially in non-scripted expressive speech also gives rise to questions regarding the perception of intonation. What are the important mechanisms and interactions involved in perceiving the multiple layers of information carried by the F0 contour?

This contribution presents a perceptual account of tonal movements in an attempt to offer possible explanations of mechanisms involved in the temporal alignment of tone and intonation. The guiding principle behind this approach is the idea that incorporating biological constraints in our understanding of the speech communication process can give us insights into why speech sounds have developed as they have and can also give us an explanatory tool to help our investigation of specific sound contrasts in different languages. This type of approach to experimental phonetics has been strongly argued for during the past twenty years particularly in the works of Lindblom [3] and Ohala [4] and has been recently represented in terms of intonation and tone by Gussenhoven [5] and Xu [6]. In a larger framework, this approach takes into consideration biological constraints in the areas of motor control (speech production and articulation), the human auditory system, (speech perception), cognition and the brain (speech planning and speech understanding), and the constraints and needs of social interaction.

The point of departure for perceptual constraints in tonal processing builds on earlier work on selective perception [7]. In this view, the biological resources available for perceptual processing of the speech signal are limited and must therefore be allocated to various tasks in a restricted manner. In terms of tonal processing the resources must be focussed selectively on the tasks of analysing the spectrum and extracting pitch from the incoming signal. The way in which these resources are allocated in time can be one of the important factors which can contribute to the shaping of patterns of tonal movement in speech. We can speculate that much of the tonal movement patterns we find in speech facilitate their perception in the same way as they are adapted to laryngeal gesture constraints in production.

2. Temporal alignment of tone

One of the more exciting and expanding issues of tonal research during the past few decades has been the question of tonal alignment. This generally refers to the time alignment of fundamental frequency excursions in relationship to segmental and syllable boundaries and has been studied from an acoustic analysis standpoint, a production perspective and a perception point of view. An important question concerning tonal alignment has been the question of variability and phonological categories. Motivated by a rule-based generative approach to intonation, the alignment of rises and falls was explored and this alignment was found to give different intonational meanings [8] and word accents [9]. These properties were first primarily investigated for pitch accent languages [10][11] but now include such issues as declarative-interrogative intonation [12] models of F0 generation and alignment [13] lexical tone alignment properties in Mandarin [14] and in the Chinese dialect of Fuzhou [15], and in Thai [16], and discourse-related intonational features [17]. In all these studies, there is a

phonological category difference which can be explained by acoustic properties of tonal alignment.

Looking at the acoustic evidence from such a wide number of languages, it is tempting to take a universal view of the importance of tonal alignment patterns for a great variety of functions. If we look at speech production, there is also growing evidence of a role of articulation constraints on tonal alignment. Based on articulation studies of maximum F0 change over time and ideas on the motor coordination of laryngeal and supralaryngeal movements, Xu [6] presents a view of production constraints for tonal alignment properties.

Finally, many perception studies concerning both tone and intonation have placed particular emphasis on the question of tonal timing. Several examples of peak shift experiments and rise-fall shift experiments are presented in [18].

A common denominator of many of the perception experiments seems to be a timing sensitivity of around 50 ms at category boundaries for both the peak shift experiments and the rise-fall experiments. What can this timing sensitivity tell us about general mechanisms of tone perception? Can we relate this sensitivity to certain segmental events that define the perception of timing? By definition, timing involves pitch movements. The tonal timing differences of more than 50 ms in perception experiments often create major tonal pattern differences on the stressed vowel or through the syllable. For example, shifting a peak can change a fall into a rise, and shifting a fall can change a low in the stressed vowel to a fall. If pitch movements are the critical perceptual cues, we can ask what is the difference, if any, between sensitivity to pitch levels and to pitch movement.

3. Perception of pitch and differential sensitivity to tonal movement

A comparison of the perception of pitch levels with pitch movement is quite closely related to the tonal concepts of level tones and contour tones in tone languages as put forth by Pike [19] and discussed by e.g. Abramson [20], Gandour [21] and Maddieson [22]. Given that the psychophysical pitch discrimination ability of humans is extremely high with difference limens on the order of 1 Hz under 250 Hz for pure tones at 40 dB SL [23] we might also expect to find languages making use of this ability and developing tonal systems in which several level tones could differ by slightly more than difference limens. Although this does not seem to be the case (microprosodic variation, intonation effects and cognitive limits probably preclude such rich tonal systems), there are tone languages in which a pair of level tones may differ in production by around 10 Hz, e.g. the mid and low tones of Thai [20] and even by less than 10 Hz in perception, e.g. perception of high and low tones in Northern Kammu [24].

The above comparison between pitch sensitivity thresholds (difference limens) and perception of level tones is fairly straightforward. A comparison between pitch sensitivity thresholds for pitch movement and perception of contour tones is more complex. Efforts have been made to compare perception of glissando tones with the perception of static tones [25] where perception of glissando tones involves an integration of the pitch movement. t'Hart [26] presents data which suggest that only differences in the amount of tonal change of more that 3 semitones are relevant for speech communication. However, the perception of tonal movement in contour tones within a given tonal system does not readily lend itself to comparison with psychophysical thresholds. For one thing, contrastive contour tones can demonstrate considerable diversification in terms of both tonal level and tonal configuration (e.g. a high falling tone compared to a low rising tone). Another consideration is that of the timing of the tonal configuration in relationship to the segmental structure of the syllable (e.g. an early fall versus a late rise).

A somewhat different approach is to vary pitch movement in relationship to segmental boundaries and ask listeners to sort and rank the stimuli in terms of pitch levels. An experiment of this type was carried out in House [18]. Figure 1 shows stylized stimuli where the carrier syllables were /amamam/ and /ama:mam/. Generally, those stimuli which fall early in the vowel (2, 4, 8, and 10) were perceived as having a lower pitch than those falling later in the vowel such as 3, 5 and 11. Furthermore, falling stimuli which are adjacent in their timing configurations early in the vowel were perceived as far from each other in pitch ranking (i.e. stimuli pairs 2/3, 4/5, 8/9, and 10/11). This critical difference does not apply to falling stimuli pairs in which the start of the fall is later in the vowel (e.g. stimuli pair 5/6). This result indicates selectivity in pitch movement perception and is consistent with earlier results presented in House [7].

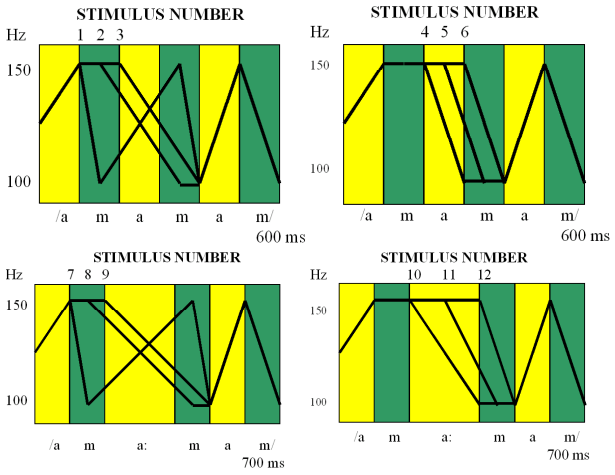

Figure 1: *Stylized F0 contours (from House 1999)*

4. Interaction between pitch and timing: question intonation and delayed peak

The signaling of interrogative mode in speech through intonation is a topic which has long attracted interest from intonation researchers. In the preceding experimental example, speech sounds were used, but the nature of the stimuli was more psychoacoustic than linguistic. A classic linguistic coding of pitch levels and movement involves the concept of delayed peak [11] which has been found to signal question intonation in many languages.

In Swedish, question intonation has been primarily described as marked by a raised topline and a widened F0 range on the focal accent [27]. An optional terminal rise has been described, but the time alignment of the focal accent rise has not generally been associated with question intonation. In recent perception studies, however, House [28], demonstrated that a raised fundamental frequency (F0) combined with a

rightwards focal peak displacement is an effective means of signaling question intonation in Swedish echo questions when the focal accent is in final position. A related production study showed that Swedish speakers produced questions with final rises in spontaneous non-scripted speech in contexts where a desire to initiate social interaction could be predicted [29].

The perception results confirmed the importance of timing where an early peak followed by a final falling contour was perceived as a statement while a late peak resulting in a rise through the final syllable was perceived as a question. An additional presence of a prefocal filled pause enhanced the percept of interrogative mode. These results demonstrated sensitivity to time alignment on the order of 25-50ms, and a perceptual interaction between pitch range and temporal alignment. Furthermore, there was a trading relationship between peak height and peak displacement so that a raised F0 had the same perceptual effect as a peak delay of 50 to 75 ms. If such a trading relationship is perceptually valid, should we then equate a delayed peak with an earlier, but higher peak? If so what is the production or perceptual mechanism behind such a coupling?

5. Biological codes and the pragmatics of human communicative interaction

The framework of biological codes for universal meanings of intonation, proposed by Gussenhoven [5] provides an elegant theoretical explanation for how delayed peak can function as the same signal as raised F0. Gussenhoven proposes three codes or biological metaphors: a frequency code, an effort code and a production code. The frequency code implies that a raised F0 is a marker of submissiveness or non-assertiveness and hence question intonation. The effort code implies that articulation effort is increased to highlight important focal information producing a higher F0. The production code associates high pitch with phrase beginnings (new topics) and low pitch with phrase endings. In this account, higher peaks take longer to reach than lower ones and thus come later in the syllable. Therefore listeners will associate a late peak with a higher pitch.

Gussenhoven's proposal for mechanisms of substitute variables whereby a peak delay can substitute for a raised peak would be an example of a mechanism shift to a higher cognitive level involving behavior. According to this argument, listeners use their knowledge that a higher peak takes longer to reach than a lower one and therefore speakers and listeners can incorporate this into a kind of cognitive pitch code. In terms of the effort code, a rise is already associated with focal accent to highlight important information. Therefore, to signal a question especially with the intention to socialize [29], the production code needs to be exploited. In the production code, high pitch is normally associated with new topics at phrase beginnings. In the case of the final rise, however, the high comes at the end of the phrase and signals the invitation to continue the social interaction.

What we see here may be evidence that perception is not limited to equating a delayed peak with a higher pitch. The rise itself may be an extra cue which is needed in certain instances being part of the production code. In the experiments presented in [28], both a high pitch and a rise are needed to unambiguously signal interrogative mode. These results can be interpreted as evidence that perception of pitch height and pitch rise are not altogether equivalent but rather can function in a complementary fashion as a simultaneous signal of two codes.

This perceptual separation of pitch levels and pitch movements may also help us explain the perception of the interaction between tone and intonation. A rough division of labor can be proposed where pitch movement has more influence on the perception of local tone identity while pitch levels are responsible for the perception of intonation. In tone languages, basic tone units, in Professor Wu's terminology, can be perceived as the presence or absence of certain language specific tonal movement. Intonation can be perceived as a sequence of pitch levels comprising different registers.

6. Tone perception, cognition and the syllable

The body of perceptual results strongly suggests that pitch alignment has a perceptual basis in the selectivity of pitch perception through the syllable. The difference between sensitivity for pitch and for tonal timing leads to a speculation that there are different perceptual mechanisms involved. The perception of level tones seems to be related to the acute sensitivity for pitch in the human auditory periphery, while sensitivity or lack of sensitivity for tonal timing seems to imply a different kind of perceptual processing, a selective processing resulting from the limited resources available for tracking pitch changes through a rapidly varying spectral configuration. In the model presented in House [7] pitch changes occurring through a rapidly changing intensity or spectral configuration will be perceived as pitch levels, while pitch changes through stable periods of intensity and spectra are perceived as pitch changes per se. Thus a falling contour early in the vowel is coded as a low tone while a slightly later falling contour is coded as a high or falling tone depending on the length of the vowel.

This somewhat simple model has been used to explain certain aspects of tonal alignment, particularly alignment at vowel onset; however, it has also been criticized [6][14] for not taking into account spectral and intensity changes frequently occurring from the syllable nucleus into the coda. Pitch movement critical for distinctive tone identification is not uncommonly found extending from the vowel through a following sonorant consonant [14] [16]. There is also mounting evidence that the onset of voicing in each syllable is a crucial point of tonal alignment (cf. House [30]). Tonal movement invariance seems to be related more to the syllable as a whole as shown recently for Thai by Ohala and Roengpitya [31]. Although this constitutes a revision of the earlier model, it is not difficult to reconcile the two. The greatest change of intensity and spectral configuration occurs at the point of syllable voicing onset. The onset of a sonorant coda entails a continuation of voicing and a lesser degree of intensity and spectral change. A certain degree of spectral change also commonly occurs during the vowel as well. Therefore from a perceptual standpoint, the most likely point of synchronization between the syllable and tonal movement would be the vowel onset or onset of voicing. Thus, a pitch movement early in the syllable would be more likely to be perceived as a pitch level than would a pitch movement later in the syllable. A final rise as a signal of interrogative mode is an example of an optimal position for movement perception. The synchronization of the tonal movement gesture with the syllable onset is consistent with the production view of Xu [6]

based on motor coordination, articulation and maximum F0 change over time.

These two types of constraints, production and perception, may also have a cognitive counterpart. With processing resources steered toward selective processing, higher-level, cognitive mechanisms may have also developed along these same lines, thus reinforcing the lower-level constraints. In this view of resource-limited processing, attention to spectral detail may be sharpened at syllable onset, while attention to pitch may be sharpened in terms of the syllable rhyme. For the processing of tonal movement and therewith the processing of tonal alignment it seems reasonable to assume such a complex perceptual mechanism involving higher order cognitive processing including short-term memory such as the precategorical acoustic storage presented in Crowder and Morton [32]. In such an acoustic storage model, pitch movement and pitch levels would be preserved in short-term memory until the tonal and intonational categories are recognized by cognitive processing. It is here on the cognitive level, that the perceptual integration of tonal movement categories and intonational levels will take place resulting in the perception of the complex tonal and intonational meaning.

7. References

[1] Wu, Z., 2000. From traditional Chinese phonology to modern speech processing – realization of tone and intonation in Standard Chinese. In *Proceedings of ICSLP 2000,* vol. 1. Beijing, China, B1-B12.

[2] Li, A., 2002. Chinese prosody and prosodic labeling of spontaneous speech. In Bel B. and Marlien I. (eds.), *Proceedings Speech Prosody 2002.* Aix-en-Provence, 39-44.

[3] Lindblom, B., 1990. On the notion of 'possible speech sound'. *Jounal of Phonetics* 18, 135-152.

[4] Ohala, J.J., 1983. Cross-language use of pitch: an ethological view. *Phonetica* 40, 1-18.

[5] Gussenhoven, C., 2002. Intonation and interpretation: phonetics and phonology. In Bel B. and Marlien I. (eds.), *Proceedings Speech Prosody 2002.* Aix-en-Provence, 47-57

[6] Xu, Y., 2002. Articulatory constraints and tonal alignment. In Bel B. and Marlien I. (eds.), *Proceedings Speech Prosody 2002.* Aix-en-Provence, 91-100.

[7] House, D., 1990. *Tonal perception in speech.* Lund: Lund University Press.

[8] 't Hart, J.; Cohen, A., 1973. Intonation by rule: a perceptual quest. *Journal of Phonetics* 1, 309-327.

[9] Bruce, G., 1977. *Swedish word accents in sentence perspective.* Lund: Gleerup.

[10] Pierrehumbert, J.B.; Steele, S.A., 1989. Categories of tonal alignment in English. *Phonetica* 46, 181-196.

[11] Ladd, D.R., 1996. *Intonation phonology*. Cambridge: Cambridge University Press.

[12] D'Imperio, M., 2001. Tonal alignment, scaling and slope in Italian question and statement tunes. In *Proceedings of Eurospeech 2001.* Aalborg, Denmark, 99-102.

[13] van Santen, J.P.H.; and Möbius, B., 2000. A quantitative model of F0 generation and alignment. In A. Botinis (ed.) *Intonation: Analysis, Modelling and Technology.* Dordrecht, The Netherlands: Kluwer Academic Publishers, 269-288.

[14] Xu, Y., 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 55, 179-203.

[15] Lin, M., 1995. Upward F0 transition in falling-falling tones and rising F0 part in falling-convex tones. *Proceedings of the Thirteenth International Congress of Phonetic Sciences '95*, vol.1. Stockholm, Sweden, 114-117.

[16] House, D.; Svantesson, J.O., 1996. Tonal timing and vowel onset characteristics in Thai. *Proceedings of the Fourth International Symposium on Languages and Linguistics,* vol. 1. Bangkok, Thailand, 104-113.

[17] Wichmann, A.; House, J.; Rietveld, T., 2000. Discourse constraints on F0 peak timing in English. In A. Botinis (ed.) *Intonation: Analysis, Modelling and Technology.* Dordrecht, The Netherlands: Kluwer Academic Publishers, 163-182.

[18] House, D., 1999. Perception of pitch and tonal timing: implications for mechanisms of tonogenesis. *Proceedings of the Fourteenth International Congress of Phonetic Sciences '99.* San Francisco, 1823-1826.

[19] Pike, K.L., 1948. *Tone Languages.* Ann Arbor: University of Michigan Press.

[20] Abramson, A.S., 1962. The vowels and tones of Standard Thai: Acoustical measurements and experiments. *International Journal of American Linguistics* 28 (No. 2 Part III).

[21] Gandour, J.T., 1978. The perception of tone. In Fromkin V.A. (ed) *Tone: a linguistic survey*. New York: Academic Press, 41-76.

[22] Maddieson, I., 1978. Universals of tone. In Greenberg (ed) *Universals of Human Language, Volume 2, Phonology*. Stanford: Stanford University Press, 335-365.

[23] Gelfand, S.A., 1981. *Hearing, an introduction to psychological and physiological acoustics*. New York: Marcel Dekker, Inc.

[24] Svantesson, J.O.; House, D., 1996. Tones and non-tones in Kammu dialects. *Proceedings of Fonetik 96, Swedish Phonetics Conference, TMH-QPSR 2/1996*, 85-87.

[25] Rossi, M., 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons del la parole. *Phonetica* 23, 1-33.

[26] 't Hart, J., 1981. Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical society of America* 69, 811-821.

[27] Gårding, E., 1979. Sentence Intonation in Swedish. *Phonetica 36*, 207-215.

[28] House, D., 2003. Perceiving question intonation: the role of pre-focal pause and delayed focal peak. *Proc 15th ICPhS.* Barcelona, 755-758

[29] House, D., 2004. Final rises in spontaneous Swedish computer-directed questions: incidence and function. To appear in *Proceedings Speech Prosody 2004*. Nara, Japan.

[30] House, D., 1996. Differential perception of tonal contours through the syllable. *Proceedings of the International Conference on Spoken Language Processing. ICSLP 96.* Philadelphia, 2048-2051.

[31] Ohala, J.J.; Roengpitya, R., 2002. Duration related phase realignment of Thai tones. *Proceedings of ICSLP 2002*. Denver, Colorado, USA, 2285-2288.

[32] Crowder, R.G.; Morton, J., 1969. Precategorical acoustic storage (PAS). *Perception and Psychophysics* 5, 365-373.