Autonomous tiers and speech perception[1]

VADIM B. KASSEVITCH
(*University of St. Petersburg*)
ANATOLY V. VENTSOV
(*Pavlov Institute of Physiology, St. Petersburg*)

1. Introduction

In numerous experiments, it has been shown fairly convincingly that, while perceiving speech, man is capable of operating with prosodic information largely independently of the information conveyed by segmentals, i.e. by vowels, consonants, and syllables. In by now classic experiments by Chistovich et al. (1965), subjects were successful in extracting rhythmic (accentual) patterns from the speech signal under white-noise or filtering masking. In their answer-sheets, the subjects adequately reproduced accentual contours of the words perceived: the number of syllables and the positions of stress were kept intact, while the words as such were very often changed. In later experiments run by one of the authors with his students and colleagues (Kassevitch et al. 1990), the same results were obtained. Under low-pass filtering experimental conditions, the subjects retained nearly all rhythmic patterns, while drastically changing the words presented to them over ear-phones. Typical were perceptual "errors" of the type *istOrija s MAshej* 'an incident with Mary' → *VictOrija plAchet* 'Victoria is crying' (capital O and A standing for stressed vowels). As one can see from the above example, the total number of syllables in the subject's response is equal to that in the stimulus phrase and the stressed positions (even the stressed vowels) are also identical. In Russian, lexical stress is not positionally fixed, this is why in a number of cases the word boundaries are found redistributed on the subjects' answer sheets, but the total number of syllables and stressed positions still tend to be retained.[2]

The autonomous nature of intonation contours can be seen from the experiments where it has been shown that, in speech perception, "the storage [of the melodic curve of the F_0] must occur independently of the lexical units at the auditory level, i.e. as an image-like representation of the contour..." (Helfrich 1984: 47).[3]

All this seems to lead to a conclusion that speech perception is a multi-channel process with different sources of information associated with specialized modules. In fact, various models has been proposed where distributed parallel processing plays a central role as is the case with connectionist models, cf. McClelland & Rumelhardt 1988 and others.

In theoretical phonology, largely independently of any experimental findings and perceptual considerations, one can trace a similar line of development, i.e. a tendency to isolate different sources of phonological information. In Ellen Broselow's words:

> "[p]erhaps the major development in post-*Sound Pattern of English* generative phonology has been the emergence of a framework in which various aspects of phonological representation (similar to those called, in other frameworks, prosodies... or long components...) are factored out of individual segments and placed on independent tiers. Among the various tiers that have been proposed is a *skeletal tier* or *timing tier*... The elements on this tier were conceived of, in most theories, as part of lexical representation, serving as the units on which higher prosodic structure was built" (Broselow 1995:175).

As is well known, the first full-blown theory of this kind was the so-called autosegmental phonology elaborated mostly by Goldsmith (1979). In Goldsmith's theory, two main tiers are introduced, one of which, roughly, is responsible for segmental representations and the other for suprasegmental ones. Special correspondence rules are introduced whose function is to relate suprasegmentals to segmentals. Perhaps, the simplest case to demonstrate how this system works can be found in tonal

languages. Skipping many details and leaving aside many formal aspects of the theory, it maintains, in fact, that two distinct strings are generated by the two tiers, one as a linearly ordered set of tones, the other as a set of syllables, also linearly organized. The two strings are conceived of as strictly independent until they are made properly interrelated due to the application of the correspondence rules referred to above.

One may be under an impression that the above model (grossly oversimplified in our presentation, but, hopefully, still recognizable) falls in nicely with the above statements tantamount to a maximal separation of segmentals from suprasegmentals in the real processes of speech perception (and, possibly, speech production alike). It could be added to what has been said above about Russian and other stress languages that, under identical experimental settings, tone languages display exactly the same behavior. Thus, in our speech-perception experiments with Chinese and Vietnamese where test words and phrases were presented to native speakers under white-noise masking (S/N ratio being equal to 0), the recognition scores for tones did not differ much from the control (about 90% on the average), while syllable recognition dramatically suffered, falling down to 50-60% (for details see Kassevitch et al. 1990).

Again one can see that both phonological formalisms and experimental results appear to point to the same direction concerned with the mutual independence of segmentals and suprasegmentals. Yet, if real speech processes and mechanisms are going to be the ultimate goal of one's study, it would certainly be desirable to specify the degree and, perhaps, the type of such an independence. Among other things, a question arises: Is it really legitimate to argue that, let us say, suprasegmental information associated with the given word can be fully specified without any reference to the segmental features of this particular word? We do not mean here the cases of compatibility vs. incompatibility of accent or tone with certain phonemes or syllables where, e.g., "heavy" and "superheavy" syllables invariably attract accent, as in Cairo Arabic (McCarthy 1979), or the case of /o/ which is incompatible with the lack of stress in standard Russian. This kind of issues is presumably accounted for by the correspondence rules. The question posited above is concerned with a (*psychological*) *reality* of a string or a set of tonal stimuli totally

deprived of their segmental basis. Notice, that segmentals, totally stripped of their prosodic (suprasegmental) distinctive features, as in our experiments with monotonized-speech perception, still retain their basic characteristics and sufficiently good recognizability, albeit not as good as in the control (Kassevitch et al. 1990).

The most natural way to test the essential independence of prosodic features isolated from their segmental basis is suggested by speech-perception experiments. Below we will describe some experiments dealing with Vietnamese speech perception.

2. Test material

As is well known, Vietnamese is a tone language where, in standard variety of the language, each syllable is marked with one of 6 tones. Phonologically, there are no atonal syllables, and no tones that would not be associated with a particular syllable ("floating tones") are attested either. In other words, in any given speech string, a one-to-one correspondence between the number of tones and that of syllables is maintained. There exist some constraints upon the combinability of certain tones and syllable types which will not interest us here.

The experimental material consisted of 12 sentences elicited from a native speaker of Vietnamese, a resident of Hanoi. Each sentence was made up of 6 one-syllable-sized words. Unfortunately, we failed to have different tones to be evenly distributed across the test sentences. The quantitative distribution of tones with reference to the positions the latter occupy in the sentences is shown in Table 1.

Table1 *Distribution of syllable tones in the stimulus set*

TONE	POSITION OF THE SYLLABLE WITHIN THE SENTENCE					
	1	2	3	4	5	6
1	8	7	4	5	6	10
2	-	3	-	3	4	1
3	1	-	-	-	-	-
4	-	-	3	1	1	1
5	2	1	5	3	1	-
6	1	1	-	-	-	-

The 12 test sentences were subdivided into three communicative subtypes, four sentences for each subtype, i.e. 4 declarative, 4 interrogative and 4 imperative sentences. All the sentences were recorded twice on audio tape.

The test sentences were then low-pass filtered at 10 kHz and digitalized at 20 kHz using 10-bit analog-to-digital converter. All the sentences were excised from the list using a digitally controlled speech wave editor on EC1036 computer. Finally, all the sentences were stored as digital stimulus files on a computer disk for further modifications and presentation during the experiment.

A special computer program made it possible to mark all the individual fundamental periods and to substitute, then, each such period with two full periods of sinusoid; the portions of the original acoustic signal associated with voiceless consonants being substituted with pauses. Thus modified signals retained the original acoustic structure, so far as their

parameters responsible for fundamental frequency, amplitude envelope, and temporal relationships were concerned. However, all the acoustic features responsible for the identification of the vowels and consonants were completely "erased".

All the resulting signals were then recorded on audio tape using a 12-bit digital-to-analog converter, the order of the 24 pseudo-sentences being randomized.

The test signals can be characterized as a kind of "humming" faithfully reproducing the melodic and other features of the original sentences but totally lacking any information about vowels and consonants. In a way, the pseudo-sentences sounded as if heard through a thick wall.

3. Subjects and experimental design

The above described signals were presented over ear-phones to 8 native speakers of Vietnamese, also residents of Hanoi or near-by provinces. The subjects were asked to listen twice to each of the 24 signals and then, after having stopped the tape-recorder, to put down any Vietnamese sentence that would tonally sound as close as possible to the "humming" they heard. The time for "inventing" responses was not limited.

In other words, we made an attempt at a literal interpretation of the autosegmental theory according to which, as stated above, a series of tonal stimuli functions as an independent entity in its own right; from this it could be inferred that such a series can be adequately perceived and recognized without any reference to its segmental (syllabic) substratum. The experimental design was partly provoked by a similar game practiced by the Japanese where a tune is played on a flute and the listeners are asked to offer a lexical match whose accentual contour would fit the tune (Rybin, personal communication).

The experimental task was considered fairly difficult by our subjects, in a number of cases the subjects tended to leave their answer-sheets blank.

3.1 Experimental results and preliminary discussion

EXPERIMENT 1

It turned out to be difficult to analyze the results, too, and for a simple reason: the number of syllables in the sentences elicited from the subjects in mere 17% of the time was equal to that of the original sentences. It is traditionally argued that Vietnamese is special in that its syllables and tones undergo only negligible reduction as distinct from other tonal languages, such as Chinese or, especially, Burmese. Yet, in real speech, Vietnamese syllables are nonetheless contracted quite perceivably or, to the contrary, lengthened due to intonational and other factors. This is why in our experiments, where subjects were denied any access to the information about the syllabic structure of the acoustically degraded messages, the syllable count proved to be hampered. In some cases, two or more "syllables" were perceived as one, whereas in other instances, on the contrary, the total number of syllables was enlarged. It was more typical of our subjects to produce sentences with a lesser number of syllables than actually uttered by the speaker.

As a result, only 17% of the answers could be analyzed with a sufficient confidence for matches vs. mismatches between the tones of the original and invented syllables. To this number, however, we could add tones of the sentence-initial and final syllables, since these syllables were anchored against, respectively, the left or the right boundary of the sentence.

The results showing percentage of the adequate tonal matches according to the position within the sentence are presented in Table 2 (the dashes correspond to tone types absent in the given position, cf. Table 1).

TONE	POSITION OF THE SYLLABLE WITHIN THE SENTENCE						
	1	2	3	4	5	6	MEAN
1	83	5	9	16	17	63	32
2	-	15	-	0	16	94	31
3	0	-	-	-	-	-	0
4	-	-	10	0	0	38	12
5	13	25	6	3	0	-	94
6	6	0	-	-	-	-	3
MEAN	255	113	83	48	83	65	-

As can be seen, some tones and some positions seem to be more favorable than the others for making adequate perceptual guesses. On the average, tones 1 and 2 can be considered better recognizable as compared to the other tones; as for the positions, the final one appears to be perceptually more salient. In the cases where the two favorable factors, i.e. tones 1 or 2 and the sentence-final position, coincide, the recognition scores become amazingly high for a non-trivial task like this, reaching 94%.

To account for the perceptual "strength" of the said tones and linear positions, one could recall that, from the point of view of intrasystemic relationships, tones 1 and 2 are UNMARKED, their acoustic

features also favoring a stable recognition. As for the sentence-final position, it is known to be salient from the point of view of the tone realization.

Yet, on the whole, one cannot argue that an ordered set (a string) of Vietnamese tones, absolutely unsupported by their respective syllables, is still adequately perceived and, hence, in any reasonable sense functionally operative. In other words, totally autonomous (independent) suprasegmentals are not fully functional.

EXPERIMENT 2

Taking into account the results of Experiment 1, we can put the following question: Given our experimental design, what kind of information would be both sufficient and necessary to make the recognition scores reasonably close to the control? Our hypothesis boils down to a suggestion that an INSERTION OF SYLLABLE BOUNDARIES might be the sought kind of information.

With this hypothesis in view, we did Experiment 2 where brief periods of silence, that is short pauses, were introduced exactly into the positions associated with consonants of the original sentences. This time all the consonants, both voiced and voiceless, were represented by the pauses whose duration was equal to the respective consonants. As a result, our experimental stimuli acquired an "intermittent" type of phonation where the inserted pauses broke down the "humming" into syllable-sized chunks.

In other words, as different from Experiment 1, in Experiment 2, the subjects were presented with the same strings of autonomous (independent) tones, BUT confined within syllable-like chunks each. Among other things (cf. below), such a structure of the acoustic signal presumably made it possible to reliably fix the initial and final points of the tonal contour, greatly contributing to the tone identification (see Kassevitch et al. 1990).

The stimuli were again presented to a team of 8 native speakers of Vietnamese, the instruction and all other conditions being exactly the same as in Experiment 1.

TONE	POSITION OF THE SYLLABLE WITHIN THE SENTENCE						
	1	2	3	4	5	6	MEAN
1	83	5	9	16	17	63	32
2	-	15	-	0	16	94	31
3	0	-	-	-	-	-	0
4	-	-	10	0	0	38	12
5	13	25	6	3	0	-	94
6	6	0	-	-	-	-	3
MEAN	255	113	83	48	83	65	-

Table 3 shows the results of the experiment. Comparing the results with those discussed above, we can see a very substantial improvement in tonal matches. The tendencies displayed in Experiment 1 still show themselves: as before, leading are tones 1 and 2 and the sentence-final position. The poorest recognition is associated with tones 3 and 6 which correlates well with their marked character, if the latter is defined in terms of featural complexity and frequency (cf. Gordina & Bystrov 1984). It seems also interesting to note that perceptual changes are typically made within the same register categories, i.e. interchanged are either the high-register tones (1, 3, and 5) or the low-register ones (2, 4, and 6).

4. General discussion and conclusion

It is self-evident that the results are not sufficient for finding out the perceptual features of Vietnamese tones with respect to the position within the sentence and other factors. To do that, one should run a special battery of experiments where all the tonal categories would be well balanced both in terms of their markedness, lexical and textual frequency , etc. Yet, the goal of the reported experiments, as stated at the outset of this paper, lies in another domain, viz.: the experiments were designed to demonstrate the LIMIT ON THE FUNCTIONAL INDEPENDENCE of suprasegmentals, here tones.

From this point of view, the experiments seem to have shown that, on the one hand, tones are relatively independent of their segmental basis, that is of "their" syllables. In fact, the very possibility of replacing tones while retaining the syllable (with a shift in semantics, of course) points to that direction. On the other hand, however, tones are found independent of INDIVIDUAL syllables, but not of the syllable as a special category. Tones, taken as such, lack any discrete character and, therefore, tend to lump together, unless demarcated due to being associated with specific syllables. As we could see, even pseudo-syllabic boundaries, making the signal discontinuous, inserting reference points for tonal domains to be defined, are found instrumental for a sharp increase in tone recognition scores.

All this means that, at least from the point of view of speech perception, tones are not absolutely independent of their segmental basis, i.e. of the syllables. The inherent link between tones and syllables is made even more visible, if one recalls the central function historically attributed to tones: tonal systems generally evolve as a device of increasing the diacritic potential of the SYLLABLES, as the latter, in tonal languages, are simply not sufficiently numerous because of heavy constraints on possible syllable structures.

The conclusion that tones are not as independent of syllables as claimed by autosegmental and other non-linear phonologists seems to hold true with respect to accent (stress), too. As mentioned earlier in this paper, perceptual errors under masking conditions typically leave untouched the location of stress AND the number of the syllables. In other words, the same inherent link between segmentals (syllables) and suprasegmentals is observed in this case as well.

One more point is to be added to the discussion of the dependence that binds suprasementals to segmentals. In Experiment 2, intonational types of the test sentences were found recognizable better as compared to the data of Experiment 1 (60% for Experiment 2 against 40% in Experiment 1). These data also speak for themselves. Although intonation contours ("intonemes") are further away from syllables than are tones (since the intonational domain is defined in terms of higher-level units such as phrases and clauses), the syllabic structure of the message turns out to be still relevant for the identification of the intonemes as well.

To sum it up, the independent tiers for segmentals and suprasegmentals are not to be doubted. Yet, when it comes to real speech, at least to speech perception, the independence of suprasegmentals from segmentals should be reasonably constrained by admitting a necessity for a robust structure of the message in terms of syllables. Such a structure being inaccessible, the chance for suprasegmentals to be adequately processed in such a "segmental vacuum" becomes strongly diminished.

This also seems to hint that the recent attempts to dispense with the syllable as a linguistic unit in its own right (cf. Dziubalska-Ko_aczyk 1996) are rather premature.

Notes

1. This paper is an enlarged and updated version of a section originaly published in Ventsov & Kassevitch 1990 (pp. 123-129).

2. As distinct from that, in the Tadjik language, where stress is always word-final, the same experimental conditions do not lead to a perceptual redistribution of word boundaries. This seems to show that primary function of lexical stress (accent) is concerned with a segmentation of the continuous auditory stream into word-sized chunks. For details see Kassevitch et al. 1990. In languages like Russian where the word-stress position is not fixed, the stress provides the hearer with an information about the NUMBER of the words rather than about the exact location of the word boundaries. Put another way, this means that, in this ("Russian") case, too, the perceptual segmentation using stress as its cue is always there, albeit WEAKLY DEFINED.

3. We won't discuss here the problem of the level of the representation of the fundamental frequency contour: it's image-like vs. "symbolic" nature.

References

Broselow, Ellen. 1995. "Skeletal positions and moras". In: Goldsmith, John (ed.) The handbook of phonological theory. Camb. (Mass.): Blackwell.

Chistovich, L. A. et al. 1965. Rech: Artikulacija i vosprijatie [Speech: Articulation and perception]. Moskva/Leningrad: Nauka.

Dziubalska-Ko_aczyk, Katarzhyna. 1996. "Natural phonology without the syllable". In: Hurch, B. & R.A.Rhodes (eds.) Natural phonology: The state of the art. Berlin/New York: Mouton de Gruyter.

Goldsmith, John. 1979. Autosegmental phonology. New York:

Gordina M. A. & I. S. Bystrov. 1984. Foneticheskij stroj vjetnamskogo jazyka [The sound pattern of Vietnamese]. Moskva: Nauka.

Helfrich, Hede. 1984. "Auditory storage of intonational contours". In: Sophia linguistica: Working papers in linguistics. No 17.

Kassevitch, V.B. et al. 1990. Udarenie i ton v jazyke i rechevoj dejatel'nosti [Accent and tone in language and speech]. Leningrad:University Press.

McCarthy, James J. 1979. "On stress and syllabification". In: Linguistic inquiry. Vol. 10 (No 3).

McClelland, J.L. & D. E. Rumelhardt (eds.) 1988. Parallel distribution processing. Vol. 2: Psychological and biological models. Camb. (Mass.): MIT Press.

Ventsov, A.V. & V.B.Kassevitch. 1990. Problemy vosprijatija rechi [Problems in speech perception]. St. Petersburg: