Transcription: what is meant by accuracy and objectivity?

Pavel Skrelin, Nina Volskaya
Department of Phonetics, Saint Petersburg State University, Russia
https://doi.org/10.36505/ExLing-2016/07/0038/000297

Abstract
The paper deals with the relationship and discrepancy between phonetic (acoustic) characteristics of the speech signal and their phonological interpretation with the aim of their reflection in segmental transcription and prosodic annotation of the speech corpora.
Key words: phonetics and phonology, transcription, speech corpora

Introduction
The presentation draws attention to the interaction between acoustic, phonetic and phonological aspects of the speech signal and their reflection in transcription. Accuracy of phonetic transcription plays an important role in the annotation of speech corpora. The requirements for precision to a great extent depend on the annotators' expertise and on what the corpus is designed for. If the corpus is to be used in TTS or ASR applications the selected phonetic signs must be as close as possible to acoustic (spectral) features of sounds analyzed in their physical boundaries. The traditional "manual" transcription is based on perception of a word or at least a syllable and represents a human model of speech perception and sound interpretation. As a result transcriptions using different methods and aimed at different applications may differ. At the same time comparison of the results of both transcription types dealt with in the presentation provides information about speech perception mechanisms on the segmental (phonetic representation of distinctive features) and suprasegmental levels (discrepancy between acoustic and perceived forms of melodic patterns).

Second paragraph text exactly the same as the first paragraph text except for the first line indentation, which is 0.6 cm for this and subsequent paragraphs (i.e. Paragraph text).

Segmental level problems
A minimal language unit for speech perception is the syllable: due to sound co-articulation distinctive features of phonemes are not limited by the boundaries of their sound realizations (allophones) proper but are

represented in their phonetic environment as well. For example, a distinctive feature of softness of the Russian plosives is actually realized in the neighboring vowels (as in the case of bilabials). Labialization of /u/ can be indicated in the preceding fricative, but may be absent from the vowel itself (as in the case of non-standard alternation of Russian phonemes /u/ - /ɨ/ we have found in CORPRESS - the Corpus of Russian Read-Aloud Speech).

This explains the use of 2 levels of representation of phonetic transcription in the corpora annotation: the first one , based on the perception of a signal fragment of a short word or syllable length (it usually corresponds to the orthoepic norm), the second one, based on the result of the perception of the sound in its physical boundaries: it reflects the sound spectral features

This method allows us to fix and describe the real situation: phoneme stream as it is perceived and interpreted by human and the same stream as it is interpreted on the basis of realized distinctive features of phonemes.

At the same time this method makes it possible to avoid solving the phonological problem, which ensues from the tensions between the abstract units (phonemes) and their material representation in the form of articulation and perception units (syllables).

Prosodic level problems

In analyzing intonation for Russian speech corpora – CORPRESS and CoRUSS (Skrelin et al. 2010; Kachkovskaia et al. 2016) – we came across situations where annotators' opinions regarding the type of a particular intonation pattern differed mostly due to the mismatch between their phonological decision and the visual acoustic representation of the intonation curve.

A few examples. In Russian, the Intonation Construction 6 (IC6) (Bryzgunova, 1970), used non-final intonation units and questions seeking repetition or clarification, is described as the (high) rising nuclear tone which levels off in the post-nuclear part. In fact, acoustically, the post-nuclear syllables form a declination line which may cover up to 4-6 semitones depending on the length of the post-nuclear part (Fig.1).

Figure 1. Schematic representation of the IC6 : nuclear syllable is marked by a bold line.

Phonologically and perceptually, though, the contour is described as "rising", and the declining part is perceptually ignored.

Another clear case for such a mismatch which complicates matters further is the use of phonetically rising-falling tone (IC3) typical for yes-no questions in Russian: though the abrupt fall on the post-nuclear part is much more prominent than in the previous case for IC6 (Fig.1) and can reach, though not necessarily, the speaker's minimum pitch level, the contour is nevertheless phonologically interpreted as rising (Fig.2).

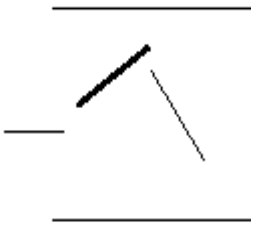

Figure 2. Schematic representation of the IC3 : nuclear syllable is marked by a bold line.

Note: For speakers of some other languages but Russian (German, English , Finnish) this contour shape is interpreted as falling. In English intonation system, for example, it belongs to the phonologically falling compex rising- falling tone , the Jackknife (O'Connor & Arnold, 1973).

This case is particularly tough both for phonological interpretation and automatic tone identification, since for any algorithm which relies on the phonetic aspect — tone-shape and F0 track only, this tone is obviously (and erroneously) falling.

Acoustically, any tone can take a number of shapes, depending on the segmental make-up of the nuclear syllable and the word itself and the location of the accented syllable proper. The case presented in Fig.3 below, shows an ambiguous situation when the tone type interpretation is unclear without postnuclear syllables, and the decision in favour of either IC6 or IC3 should be taken with other prosodic parameters in consideration, namely, the nuclear syllable duration, which is normally longer in IC 6.

Figure 3. Schematic representation of the IC6 and IC3with nuclear syllable in the final position.

Conclusion

In real speech situation the distictive features cruicial for the phonological decision-taking may not be present in the sound itself (which may be absent altogether) but reflected in its right or /and left neighbours. This poses the problem of formal represenation of the sound stream itself in automatic interpretation (recognition) which is based on acoustic parameters of segments or F0 curves. As long as we do not exactly know how the speech signal characteristics which a person uses for phonological interpretation correlalte with its objective evidence we need to use two ways of formal representation (transcription): objective and abstract.

References

Bryzgunova E. A. 1980. Intonation [intonacija], in: Russian Grammar, N. Shvedova, Ed. Moscow: Nauka, vol. 2, pp. 96 – 122.

Kachkovskaia T., Kocharov D., Skrelin P., Volskaya N. 2016. CoRuSS - a new prosodically annotated corpus of Russian spontaneous speech. in: Proceedings of LREC 2016.

O'Connor J.D., Arnold G.F. 1973 Intonation of Colloquial English. Longman, London.

Skrelin P., Volskaya N., Kocharov D., Evgrafova K., Glotova O., Evdokimova V. 2010. CORPRES - Corpus of Russian Professionally Read Speech. in: Text, Speech and Dialogue, ser. Lecture Notes in Computer Science, P. Sojka, A. Hor´ak, I. Kopecek, and K. Pala, Eds. Springer Berlin Heidelberg, 2010, no. 6231, pp. 392–399.