Speech Tempo Control in Automatic Speech Synthesis

O.F.Krivnova
Lomonosov Moscow State University

Abstract

Any speech synthesizer, even of the simplest kind, includes the option of imitating the general tempo of speaking for an utterance to be synthesized. Unfortunately, the developers tend to let a user himself choose the necessary tempo parameters. If the tempo features are assigned and realized automatically by the rules, it often remains unclear to what extent these rules take into account the phonetic data concerning realization and perception of tempo differences in natural speech. The present report contains recommendations for tempo parameters control obtained for the Russian language as a result of special experiments including both the acoustic analysis of a representative speech material (about 1 hour of duration) and the study of perception of tempo contrasts by Russian native speakers. The reported recommendations have been practically tested in a synthesizer of the Russian speech developed by the Speech group of Philological Faculty at Moscow State University

.

1. Introduction

The general tempo of an utterance is usually defined as a mean speed of its articulation. The traditional use of this parameter in phonetic studies and systems of automatic speech synthesis is based on the following ideas. On the one hand, any person has his/her individual tempo of articulation, this being a feature of the speaker's speech portrait. At the same time, as has been shown experimentally [1; 2; 3], a man can control the change of the utterance's tempo independently of other speech parameters, both in speech production and perception, by using fairly stable temporal standards of various tempo categories. Traditionally three categories are distinguished: fast, medium or normal and slow tempo of speech. On the other hand, there are a number of reasons to consider the differences in the general tempo of utterances to be functionally connected with the communicative importance of information comprised in the utterance, as well as with the psycho-emotional state of a speaker at the moment of communication etc. [4].

2. The problems of tempo control in TTS systems

At TTS-synthesis the reproduction of human ability to control the speech tempo triggers two independent but closely connected problems, which are provided by the existing phonetic studies in a different degree. The first problem consists in both defining the type and the size of a text fragment, within which a certain tempo category is realized. Also it is necessary to choose a definite tempo category from the resolved set.

When synthesizing a narrative text the definite tempo feature is usually assigned to a whole sentence (a part of the text from a point up to a point) or to a separate intonational phrase (a minimal sense-bearing intonation unit of the text). In the latter case the synthesizer should have the rules which automatically realize intonational (prosodic) phrasing of the text. At synthesis of a dialogue the tempo feature is usually attributed to a separate retort.

As far as the choice of concrete tempo features for "pronouncing" separate text fragments, this problem in an automatic mode, i.e. based on the analysis of the text itself, is not solved yet. And more so it is not quite clear *what* exactly should be searched for in the text for assigning the tempo feature needed. There are very few data concerning the functional use of tempo contrasts in real speech. When solving practical synthesis tasks the text is either tagged manually from the standpoint of tempo differences between its fragments, or is synthesized at one and the same tempo, usually medium by default.

The second range of problems is connected with phonetic questions proper, that is, with developing the time parameterization rules, by which the tempo feature chosen is realized in the duration patterns of phonetic units comprising the utterance. There are certain difficulties here too, though the phonetic experience is broader and more diversified than in the case of the problems of the first type discussed earlier. The principal cause of timing difficulties is the absence of clearly stated requirements for the element of the speech flow that serves as a bearer of tempo distinctions.

It is natural to use the term 'a tempo-difference bearer' (TDB) for such elements of speech, whose temporal characteristics or frequency in time can be used as a measure of the tempo contrasts. Different phonetic studies use the following units as a bearer of tempo properties: sound segment, syllable, morpheme, word (either grammatical or phonetic). Such a variety doesn't help in understanding the nature of tempo and leads to complications in comparing the results obtained by different researchers. That is why we can't but consider fair the opinion [5], that a proved choice of a TDB-element both for phonetic studies and applied research is not possible without taking into account the speech behavior of native speakers, their perception of tempo distinctions. In a number of studies there have been made an attempt to define the most important properties which the ideal speech element – TDB should possess:

a) The phonetic dimension of the element - TDB should be smaller than that of a speech unit which tempo is to be measured;
b) TDB should match such segments of a speech signal, boundaries and number of which can be defined basing on the properties of the signal itself without use of the information about sound (phonemic) structure of a speech unit, i.e. without its preliminary phonemic identification.

It has been confirmed experimentally [3], that a listener is able, being given the only acoustic information, to divide the speech signal into segments corresponding to vocalic and non-vocalic parts and to define the duration of these segments. These facts lead to the conclusion that the most probable candidates for the role of TDB are a syllable (as defined traditionally) and a vocalic cycle (a fragment of a sound sequence from the beginning of a vowel to the beginning of the next one with an obligatory presence of at least one consonant between them, i.e. a sequence **VCn (VC)**, the duration of which defines the period of vowels' succession in the utterance). The experiments described in [5] have delivered the data proving that the initial physical information for tempo perception is the duration of the vocalic cycle **VC**. These experiments also show that the speech utterance on the whole may be characterized by the integral tempo feature based on the mean arithmetic duration of VC-intervals forming it.

Apart from a reasonable choice of a phonetic unit - TDB the anthropomorphic strategy of tempo control for speech synthesis should take into account the data of absolute (categorial) and relative (differential) thresholds of tempo perception. Unfortunately, this problem remains insufficiently explored. The data of differential sensitivity of hearing to periodicity of sound events, obtained in psycho-acoustic experiments, give very approximate knowledge of possible values of hearing threshold of general tempo changes. According to these data, in the range of tempos typical for speech the relative differential periodicity threshold does not exceed 6% [6].

However, the application of a certain rhythm to a sound sequence raises the threshold of detection of changes in periodicity up to 20-22% [7]. Besides, it is unclear how the threshold of tempo distinctions (when the utterances are compared by pairs) correlates to the categorial tempo evaluation. It also remains unknown, how the categorial tempo estimates are affected by those non-tempo linguistic factors, on which the mean value of the vowel succession period (or mean syllable duration) depends (for instance, such factors as the sound composition of an utterance, the ratio of stressed and unstressed vowels, rhythmical structure etc).

The detailed study of the listed phonetic problems is a special challenge. The present report describes the results of perceptive experiments conducted to obtain at least preliminary answers to the questions mentioned above.

3. Experiments

In the experiments described below we were mostly interested in the following peculiarities of the perception of tempo differences between intonational phrases (IP) of a connected text:

1. What is the degree of coordination of perceptive tempo estimations given by different subjects for the same IP when distinguishing between three standard tempo categories: fast, normal (medium), slow;
2. With which of the temporal parameters - mean duration of a syllable or a vocalic cycle - the perceptive tempo estimates are correlated more;
3. What are the absolute tempo thresholds corresponding to different tempo categories and do they depend on the phonetic structure of the utterance (sound composition, rhythmic pattern);
4. What are the relative tempo thresholds in the domain of the observed values of the mean vocalic cycle duration for the utterances of the same or similar phonetic structure;
5. Do the tempo contrasts affect the consonant and vowel segments of vocalic cycle in the same way?

Our research [8; 9] was based on a scientific linguistic text, which was read in normal tempo by a male speaker of the Russian Moscow standard pronunciation, a linguist. The text was recorded in a studio with the use of a high-quality tape recorder. The total duration of the recorded material is 45 minutes. The recording was later digitized by a computer (SR=11025 Hz, 8 bit AR).

The perception tests were run under the following conditions. The subjects (10 people) were to listen to 2-3-word intonational phrases (351 units) randomly cut out of the recorded text. Successive IPs were played at standard pause intervals of 6-7 seconds. To form a general idea of the individual tempo of the speaker, at the beginning of the test the subjects were to listen to a rather large text fragment (total duration about 3 minutes) . When listening to each separate IP the subjects had to evaluate the relationship between its general tempo and the individual tempo of the speaker. Three estimates were supposed: they coincide, i.e. the given IP is read at a normal tempo (NT), it is read at a fast tempo (FT) or slow tempo (ST).

The analysis of the perception evaluations obtained leads to the following conclusions:

- In the vast majority of cases the general tempo of a given IP is evaluated by different subjects in the same way, i.e. has 70% of matches. 87.5% of given utterances turn out to have received identical perceptive estimates. The analysis of correlation between perceptive evaluations and the mean duration of presumable TDB- elements shows that the mean duration of the vocalic cycle is a more adequate measure of tempo contrasts than the mean syllable duration, though the differences in correlation coefficients are not considerable.
- If we choose the mean duration of the vocalic cycle T(VC) as a physical measure of tempo, the change of categorial tempo estimates lies in the intervals of **130-150 ms** for the **FT → NT** transition and **210-230 ms** for the **NT → ST** transition. The tempo of the utterances with T(VC) belonging to the intermediate zone **150-210 ms** is estimated as normal (medium) in no less than 70% of cases. An additional analysis of the perception estimates shows that the temporal characteristics of the boundary transitions (absolute tempo thresholds) do not depend on the phonetic structure of the analyzed utterances.

In connection with the results obtained it should be noted that the values of T(VC) corresponding to the categorial tempo thresholds, are the integer products of the minimal duration of a syllable needed for its detection in the speech signal. According to [3], this detection threshold Tthr equals to **65-70** ms. This results in the following ratios of the estimates we have obtained: T(VC) for the (FT → NT) transition = 2 Tthr; T(VC) for the (NT → ST) transition = 3 Tthr. It can be assumed that the threshold duration of the VC detection is a natural measure of the perception estimates of tempo contrasts.

The analysis of the change of tempo estimates depending on the deviation of the mean T (VC) observed for the given IP from the T(VC), typical of the utterances with the similar phonetic structure whose tempo is evaluated as ‘normal’, shows that the relative tempo thresholds are

characterized by asymmetry. The threshold values lie in the semi-interval **(-20, -15] %** for the transition from normal to fast tempo and **(30, 35] %** for the transition from normal to slow tempo of speaking. Respective tempo control coefficients can be evaluated as (0.80 – 0.85) for the increasing tempo and (1.30-1.35) for the decreasing tempo.

Tempo distinctions affect consonant and vowel segments of the utterance in a different way: when the tempo increases the consonant part of VC is more considerably reduced than the vocalic one – 14% and 10% respectively. When the tempo decreases the spreading of the vowel part is more pronounced. The subtler features of tempo control, for instance, the probability of different kind of influence of tempo changes on the smaller parts of sound segments [10] demand the separate studying.

The results discussed above have been used in the TTS synthesizer of Russian speech developed by the Speech group of Philological Faculty at Moscow State University. In this system there is an option of separate modification of consonants' and vowels' durations when controlling the tempo of speaking. The normal tempo is not specifically adjusted, as the timing rules for sound segments depending on various phonetic factors are normal tempo-oriented by default. As far as fast and slow tempos are concerned, the automatic control mode employs default threshold values and coefficients reported above. The manual control mode includes the option of using any relative coefficients of the tempo change in respect to the normal tempo or the other tempo feature assigned by default.

References

1. *Agafonova L.S., Bondarko L.V., Verbitskaya L.A. et al.* Some characteristics of Russian speech depending on different tempo of pronunciation // Hearing and Speech in the norm and pathology. 1974. V.1 (in Russian)
2. *Chistovich L.A., Kozhevnikov V.A. et al.* Speech: Articulation and Perception. M.-L ., 1965 (in Russian)
3. *Chistovich L.A., Ventsov A.V. et al.* Physiology of speech. Perception of speech by a human. M.-L., 1976 (in Russian)
4. *Theplitis L.K.* Analysis of speech intonation. Riga, 1974 (in Russian)
5. *Ventsov A.V.* Speech tempo and some peculiarities of its perception. // Sensory systems. L., 1977 (in Russian)
6. *Ventsov A.V., Malinnikova T.G.* Modeling of the subjective mechanism of duration comparison // Research of models of speech production and perception. L., 1981 (in Russian)
7. *Krylov I.N.* To a question on regulation of rate of simple rhythmic movements // Movements' control. L., 1970 (in Russian)
8. *Krivnova O.F.* Perception of the integral speech tempo of a syntagma // Proc. of the Seminar
9. "ARSO14 ". Kaunas, 1986 (in Russian)
10. *Krivnova O.F.* Duration of a vocalic cycle and perceptive tempo thresholds // Proc. of the Seminar "ARSO-15 ". Tallinn, 1989 (in Russian)
11. *Zinder L.R.* Influence of speech tempo on articulation of some sounds // Philological sciences. 1964. V.69 (in Russian)