STUDIES OF PHONETIC FUNCTION APPLICABILITY TO AUTOMATIC QUALITY TESTING OF SPEECH PROCESSING SYSTEMS AND TRANSMISSION CHANNELS

V.N. SOBOLEV

ABSTRACT

The quality of a speech channel or codes is tested by comparing the phonetic functions at its input and output. The results of experimental verification of this method are presented, its hardware implementation is discussed.

INTRODUCTION

An objective method of transmission quality estimation is essential in designing new speech processing equipment, such as vocoders and speech waveform coders, as well as in maintenance of existing communication channels. At present, this is performed by averaging the subjective responses of a large number of listeners. Such statistics yield consistently accurate results, but are time- and labor-consuming. An objective quality measure would facilitate detection of distortion causes and thus improve the efficiency of designing new speech coders. In maintenance this would provide automatic monitoring of channel quality for timely replacement or adjustment of faulty units.

The problem of developing, study and use of methods of objective speech transmission quality measurements has been addressed by a number of authors. The methods discussed include correlation techniques [1,2], segment and subjective signal-to-noise ratio [3,4], isopreference method [5], various physical methods (e.g., as in [6]), log likelihood ratio measure [7], and the Itakura-Saito measures [8]. However, some of them are rather complicated in actual practice, others are insensitive to certain types of distortion, and yet others are not always adequately accurate. Thus, the problem of automated objective quality testing retains its urgency and further research for new approaches is justified. Presented here is an attempt to introduce a unique transmission quality index based on comparing phonetic function at a system's input and output.

THE METHOD

The phonetic function

\[ P(\omega, t) = \frac{1}{S(\omega, t-t)} \cdot \log \frac{S(\omega, t-t)}{S(\omega, t)} \]

where \( S(\omega, t) \) is the modulus of the speech signal's \( f(\ell) \) short-term spectrum, was first introduced by A.A. Zheglov [9] and N successfully used for phoneme recognition [10]. The feasibility of evaluating speech transmission quality with this function has been demonstrated [11]. The logarithmic term in this equation describes the increment in information amount during the time interval \( T \) at a moment of time and frequency \( \omega \). This phonetic function is based on the Weber-Fechner psychophysiological law and takes human ear adaptation effects into account, i.e., the logarithmic relation between sound perception and intensity of auditory analyser excitation along with the human ear tending to perceive subsequent sounds against a background of impressions from preceding sounds.

The phonetic function describes only the dynamics of speech signal spectrum variations in the time domain.

Comparing phonetic functions at a code input and output yields the following equation to describe a criterion of speech transmission quality:

\[ q = \frac{\int_{\omega_1}^{\omega_2} P(\omega, t) R(\omega, t) \cdot d\omega}{\int_{\omega_1}^{\omega_2} R(\omega, t) \cdot d\omega} \]

where \( w(\omega) \) is the weight function to deform the frequency axis according to the Koenig scale; \( \omega_1 \) and \( \omega_2 \) are the lower and upper frequencies of the band under study; \( t_1 \) and \( t_2 \) limit the comparison time interval; and symbol \( R \) depicts the comparison operation. Speech signals differ more from the other in spectrum distribution time variations, rather than in spectrum themselves and therefore comparing signals by their phonetic functions is both justified and feasible.

EXPERIMENTAL VERIFICATION

The experiments included comparing the phonetic functions at inputs and outputs of speech waveform coders and comparing these results with those of articulation tests. EM and ECM devices were tested at various transmission rates with syllabic intelligibility ranging from 30% to 60%. Measurements were computerized, with the test signal in the form of a tape record of linguistically balanced speech of 42-second duration, from four different dictors. The computer input signals were the logs of the speech signals short-term spectra, \( \log S(\omega, t) \) and \( \log S(\omega, t-t) \), from the outputs of band-pass analysers, rather than the initial speech signals \( f(\ell) \) and \( f(\ell-t) \). The center frequencies of the 16 band-pass filters range from 30 to 7 kilohertz and are distributed according to Koenig's scale, thus realizing the \( w(\omega) \) weight function. The phonetic functions were compiled by recursion formulas

\[ P(\omega_1, n-t)=\frac{2}{\pi t} \log S(\omega_1, n-t) - H(\omega_1, n-t) \]

\[ H(\omega_1, n) = \exp \left(-\frac{\alpha}{T} \right) \cdot H(\omega_1, n-1) \]

along with the quality index \( q \), with various comparison techniques, the best of

\[ q = \frac{\int_{\omega_1}^{\omega_2} P(\omega, t) \cdot d\omega}{\int_{\omega_1}^{\omega_2} R(\omega, t) \cdot d\omega} \]

Fig. 1. Quality index \( q \) vs. syllabic intelligibility \( s \) (with different methods of comparing phonetic functions)

Fig. 2. Individual channel quality indices \( q_i(s) \) (broken lines) and the averaged function \( q(a) \) (solid curve)
which turned out to be:

\[ P(\beta) = P(\beta - P) \]

for the digamma function with the argument \( \omega_1 \) and \( \omega_2 \), where \( P \) is the asymptotic function and \( \beta \) is the parameter of the function.

These computations resulted in monotonic functions which relate the quality index \( q \) to the syllabic intelligibility \( s \); examples are shown in Fig. 1. The highest sensitivity was obtained with the second comparison technique. Relations

\[ q_1 = \frac{\omega_1 \cdot s}{\omega_1 + \omega_2} \quad \text{for individual band-pass filters are shown in Fig. 2 in broken lines along with the result relation} \]

\[ q = \frac{16}{\sqrt{16}} q_1 \]

shown as a bold curve, confirming the expedience of covering a wide frequency band. Fig. 3 shows the average slope of \( q(s) \) curves, calculated by \( \frac{\omega_1 \cdot s}{\omega_1 + \omega_2} \) vs. the averaging time constant \( T \). The highest sensitivity measurement was obtained at \( T \) from 200 to 1000 ms. The effect of a time shift between \( f(t) \) and \( f(t) \) was studied, with measurements showing the nonmonotonicity of \( q(s) \) curves being maintained at time shifts up to 50 ms, but at steadily decreasing slope angles and accompanied by an upward displacement of the curve. This means that the presence of a time shift between the signals being compared impairs the measurement sensitivity and leads to underestimating the measured quality index. Measured absolute \( s \) and relative \( \delta \) errors of automatic syllabic intelligibility \( s \) measurements are shown in Fig. 4 as functions of the time shift \( \delta \) for systems of various transmission quality.

These measurements are complicated by the inverse proportionality between \( q \) and \( s \); a modified criterion \( Q = (r/q)^{-1} \) (where \( r \) and \( Y \) are empirical coefficients) proves to be more convenient. When using the second comparison technique, the best fit was obtained by setting \( r = 0.71 \) and \( Y = 0.13 \) for time constants \( T \) from 500 to 1000 ms; \( q(s) \) under these conditions is as shown in Fig. 5, which confirms the correspondence between \( q \) and \( s \).

REFERENCES