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

V.N. SOBOLEV

All-Union Correspondence Electrotechnical Institute of Communications, Moscow, USSR 123855

ABSTRACT

The quality of a speech channel or codec 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 responces 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 ratios [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

$$P(\omega, t) = \int_0^{\infty} \exp(-\tau/T) \cdot \log \frac{S(\omega, t)}{S(\omega, t-\tau)} d\tau$$

where $S(\omega,t)$ is the modulus of the speech signal's f(t) short-term spectrum, was first introduced by A.A.Pirogov [9] and is successfully used for phonem 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 τ at a t moment of time and frequency ω. 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 audial analyzer excitation along with the human ear tending to percieve subsequent sounds against a background of impressions from preceeding sounds. The phonetic function describes only the dynamics of speech signal spectrum variations in the time domain.

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

$$q = \int_{\omega_1}^{\omega_2} w(\omega) \int_{t_1}^{t_2} P(\omega,t) \circledast \tilde{P}(\omega,t) \cdot dt \cdot d\omega,$$

where: $w(\omega)$ is the weight function to deform the frequency axis according to the Koenig scale; ω_1 and ω_2 are the lower and upper frequencies of the band under study; t_1 and t_2 limit the comparison time interval; and symbol $\circledast$ depicts the comparison operation. Speech signals differ more from one another in spectrum distribution time variations, rather than in spectra 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 codecs and comparing these results with those of articulation tests. DM and PCM devices were tested at various transmission rates with syllabic intelligibility ranging from 30 % to 80 %. Measurements were computerized, with the test signal in the form of a tape record of phonetically balanced speech of 42 seconds duration, from four different dictors. The computer input signals were the logs of the speech signals short-term spectra, $\log S(\omega_i,t)$ and $\log \tilde{S}(\omega_i,t)$, from the outputs of band-pass analyzers, rather than the initial speech signals f(t) and $\tilde{f}(t)$. The center frequencies of the 16 band-pass filters range from DC to 7 kHz and are distributed according to Koenig's scale, thus realizing the $w(\omega)$ weight function. The phonetic functions were computed by recursion formulas

$$P(\omega_i, n\cdot\Delta t) = (T/\Delta t)\log S(\omega_i, n\cdot\Delta t) - H(\omega_i, n\cdot\Delta t)$$

$$H(\omega_i, n\cdot\Delta t) = \exp(-\Delta t/T)\cdot H(\omega_i, (n-1)\cdot\Delta t) + \log S(\omega_i, n\cdot\Delta t)$$

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

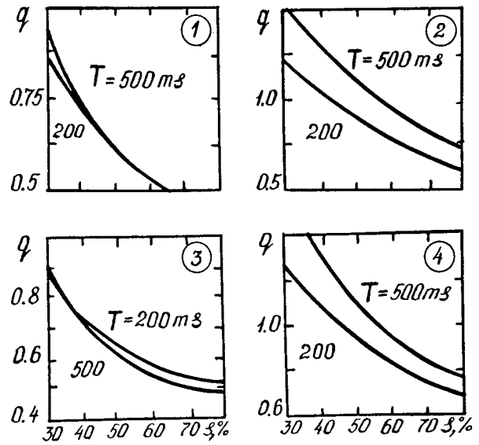

Fig. 1. Quality index q vs. syllabic intelligibility s (with different methods of comparing phonetic functions)

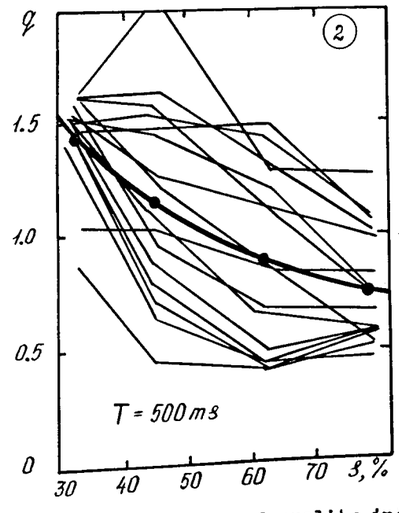

Fig. 2. Individual channel quality indices $q_i(s)$ (broken lines) and the averaged function q(s) (solid curve)

which turned out to be:

$$P①\tilde{P} = |P - \tilde{P}|; \quad P②\tilde{P} = (P - \tilde{P})^2;$$

$$P③\tilde{P} = \frac{|P - \tilde{P}|}{\max(|P|,|\tilde{P}|)}; \quad P④\tilde{P} = \frac{(P - \tilde{P})^2}{\max(P^2, \tilde{P}^2)}.$$

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

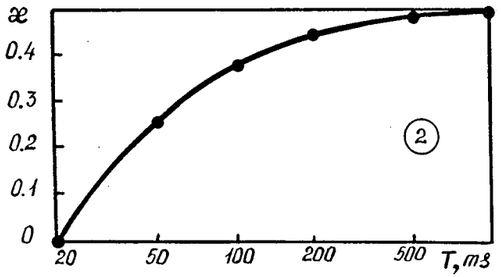

Fig. 3. Average slope of q(s) curves vs. averaging time T

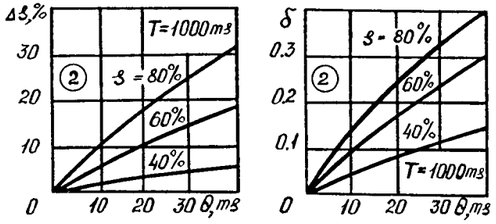

Fig. 4. Absolute (Δs) and relative (δ) syllabic intelligibility measurement error vs. time shift between signals being compared

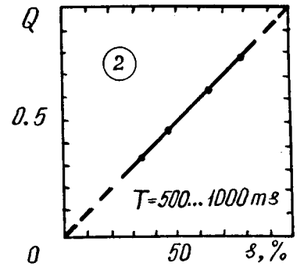

Fig. 5. Quality index Q and syllabic intelligibility s

sensitivity was obtained with the second comparison technique. Relations

$$q_i = \int_{t_1}^{t_2} P(\omega_i,t) \quad ② \quad \tilde{P}(\omega_i,t)dt$$

for individual band-pass filters are shown in Fig.2 in broken lines along with the resulting relation

$$q = \sum_{i=1}^{16} q_i$$

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 $æ = (q(33) - q(78))/q(33)$, vs. the averaging time constant T. The highest measurement sensitivity was obtained at T from 200 to 1000 ms.

The effect of a time shift between f(t) and $\tilde{f}(t)$ was studied, with measurements showing the monotonicity 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 curves. This means that the presence of a time shift between the signals being compared impairs the measurement sensitivity and leads to underrating the measured quality index. Measured absolute Δs and relative δ errors of automatic syllabic intelligibility s measurements are shown in Fig.4 as functions of the time shift θ for systems of various transmission quality.

These measurements are complicated by the inverse proportionality between q and s; a modified criterion $Q = (r/q) - \varphi$ (r and φ 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 φ = 0.13 for time constants T from 500 to 1000 ms; Q(s) under these conditions is as shown in Fig. 5, which confirmes the correspondence between Q and s.

HARDWARE IMPLEMENTATION

From the above it follows, that automatic measurements of speech transmission quality can be provided by arrangement shown in Fig. 6, which functions as follows. The log of the output signals from two band-pass analyzers are fed to the inputs of channel integrators. In each of the channel adders the integrator output signal is subtracted from its input signal, thus producing a signal proportional to the log of the ratio of the short-term spectrum at this time moment to the integral of this spectrum over the preceeding

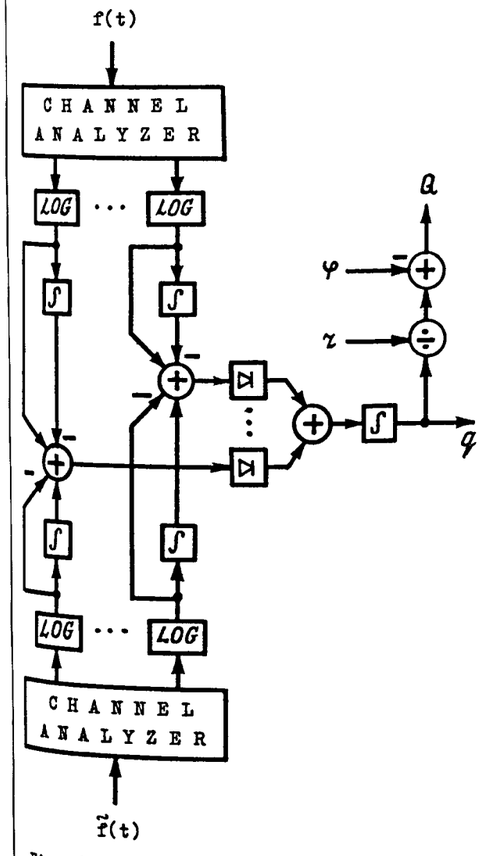

Fig. 6. Skeleton diagram of measurement circuit arrangement

time interval. The set of such difference signals in all channels represents the speech signal phonetic function. Since the difference signals from the upper and lower sections of this circuit arrive at the channel adders in antiphase, the output of these adders is the difference between the phonetic functions of speech signals f(t) and $\tilde{f}(t)$. The channel adder output signals are squared, summed and averaged in a group integrator over a significantly longer time interval. The output voltage, q, of this group integrator is then converted into Q. This latter quantity is the measure of transmission quality via a codec or channel under measurement.

REFERENCES

1. М.А.Сапожков. Акустический журнал, 1956, вып. 3, с. 279 - 284.
2. Т.Е.Зайцев. Электросвязь, 1958, № 10, с. 38 - 46.
3. C.Scagliola. BSTJ, v.58, Jul.-Aug. 1979, pp.1369-1394.
4. M.Nakatsui, P.Mermelstein.JASA, v.72, No.4 (1982), pp.1136-1144.
5. W.A.Munson, J.E.Karlin. JASA, v.34 (1962), pp.762-774.
6. H.J.M.Steeneken, T.Houtgast. JASA, v.67 (1980), pp.318-326.
7. R.E.Crochiere, J.M.Tribolet, L.R.Rabiner IEEE Trans.Acous., Speech Signal Processing, v.ASSP-28, No.3, June 1980, pp.318-323
8.B.-H.Juang. On using the Itakura-Saito measures for speech coder performance evaluation.- AT&T Bell Lab.Techn.J., v.63, No.8, Pt.1, 1984, pp. 1477-1498.
9. А.А.Пирогов. Электросвязь, 1967, № 5, с. 24 - 31.
10. Вокодерная телефония / под ред. А.А. Пирогова - М.: Связь, 1974.
11. В.Н.Соболев, Г.В.Титова. Автоматическое распознавание слуховых образов. Тезисы докладов Всесоюзного семинара АРСО-8, с. 87 - 89, 1974.