INFORMATICA, 1999, Vol. 10, No. 2, 171–188

Noisy Speech Intelligibility Enhancement

Kazys KAZLAUSKAS
Institute of Mathematics and Informatics, Vilnius Pedagogical University
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: kazlausk@ktl.mii.lt

Received: March 1999

Abstract. This paper addresses the study of the speech intelligibility enhancement. The speech model, noise sources, perceptual aspects of speech, and performance evaluation are reviewed. The intelligibility enhancement system based on spectral subtraction technique is investigated. Spectral density estimation device based on the algorithm of smoothed periodograms is analysed. Determination of the silence intervals, efficiency of the silence intervals determination, and signal to noise ratio evaluation are discussed. Speech intelligibility enhancement device is described.

Key words: noisy speech, intelligibility enhancement.

1. Introduction

The objective of the speech enhancement may be to improve the quality, to increase intelligibility, to reduce listener fatique, ect. (Thomas and Ravindran, 1974; Curtis and Neiderjohn, 1978; Sambur, 1978). A speech communication system may introduce a low-amplitude long time delay echo. Another example is the communication between a pilot and an air traffic control tower. In this case the speech is degradated by background noise. Of central importance is the intelligibility of the speech and it would generally be acceptable to sacrifice quality if the intelligibility could be improved. The speech enhancement problem covers a broad spectrum of constraints, applications and issues. Environments in which an additive background signal has been introduced are common. The background may be noise-like such as in aircraft, street noise, etc., or may be speech-like such as an environment with competing speakers. Other examples in which the need for speech enhancement arises include correcting for reverberation, correcting for the distortion of the speech of under-water divers breathing a helium-oxygen mixture. The problem and techniques vary, depending on the availability of the signals or information. For example, for enhancement of speech in an aircraft a separate microphone can be used to monitor the background noise so that the characteristics of the noise can be used to adjust or adapt the enhancement system. We assume that the only signal available is the degradated speech and that the noise does not depend on the original speech.

Speech is a special subclass of audio signals and there are models in terms of which the speech waveform can be described. The more specifically we attempt to model the speech signal, the more potential for separating it from the background noise. On the other hand, the more we assume about the speech the more sensitive the enhancement system will be to inaccuracies from these assumptions. Thus incorporating assumptions and information about the speech signal represents tradeoffs which are reflected in the various systems.

Another important consideration in speech enhancement stems from the fact that the criteria for enhancement relate to an evaluation by a listener. In different contexts the criteria for evaluation may differ depending on whether quality, intelligibility, or some other attribute is the most important. Speech enhancement must take into account aspects of human perception. Some systems are motivated by perceptual considerations, others rely more on mathematical criteria. The mathematical criteria must be consistent with human perception (Lim and Oppenheim, 1979).

2. Speech Model and Perceptual Aspects

2.1. *Speech Model*

Speech is generated by exciting of noisy speech the vocal tract by pulses of air released through the vocal cords for voiced sounds, or by turbulance for unvoiced sounds. Thus model for speech production consists of a linear system, representing the vocal tract, driven by an excitation function which is a periodic pulse train for voiced sounds and wide-band noise for unvoiced sounds. Many of the techniques for speech enhancement are based on the representation of the speech signal as a stochastic process. That is more appropriate in the case of unvoiced sounds for which the vocal tract is driven by wide-band noise. The vocal tract changes shape as different sounds are generated and this is reflected in a time-varying transfer function of the linear system. It is responsible to represent the linear system as a slowly varying linear system so that on a short-time basis it is approximated as stationary. This simplified model for speech production has been successful in a variety of engineering contexts including speech enhancement, synthesis, and bandwidth compression (Fant, 1970; Flanagan, 1972; Rabiner and Scafer, 1978).

2.2. *Noise Sources*

Noise induced by the environment or the transmission channels can be linear in the power spectrum domain or linear in the log spectral or cepstral domain or non-linear in both domains. Environment noises are usually additive. Depending on the environment, such assumptions may or may not hold. Another assumption, often made is that the noise is stationary and uncorrelated with the speech signal. The long term stationarity of the noise excludes distortions occurring frequently in office environments such as door slams and speaker induced noise. Noise sources can be classified in various types according to the task and the context in which the conversation occurs.

Office noise is induced by a variety of causes such as typewriters, computers, fluorescent lamps, chair movements and etc. This type of noise is additive and is characterized by energy concentrations over certain portions of the spectrum. The sources of channel distortion over telephone lines can be separated in various categories such as burst or impulse noise, hum, additive stationary noise, unknown channel gain and phase response, echo, intermodulation distortion, etc. Some of these distortions are additive in the log spectral domain.

In rooms which have hard reflecting surfaces, there is a significant reverberant field, which creates a *room acoustics*. A sound spoken in a room is prolonged, with a more or less logarithmic decay, so that it is present to mask subsequent sounds. In telephone speech there is often reverberation when the microphone is placed too far from the talker. The type of noise induced is a convolution noise.

Transportation noise includes noise occurring in vehicles running on roads such as cars, on railways such as trains, and in the air such as aircraft. Depending on the type of transportation, noise will be caused by different factors. For cars it is mainly due to engine, for trains it is due to the locomotive, while for aircraft the main sources are aerodynamic noise, vibrations, and compressor.

Military noises are very diverse in their nature. They are generally additive in the spectrum domain. The military noises are wide-band and have strong frequency components at certain frequencies (Janqua and Haton, 1996).

2.3. *Perceptual Aspects of Speech*

The effects of noise on people are numerous. Speech interference can result in annoyance and tiredness. Effect of noise on man is its incidence on communication (Abel and Alberti, 1982). At a fairly defined level, noise will mask the sound communicated. It is known that the linear prediction analysis is very sensitive to the noise effects. In many practical applications signal to noise ratio (SNR) from 0 to +25 dB is desirable. To attenuate noise distortions, noise canceling, close-talking together with speech enhancement and noise compensation techniques, have been used. There are a number of aspects which play an important role in enhancement systems. For example, consonants are known to be important in the intelligibility of speech even though they represent a relatively small fraction of the signal energy. It is understood that the short-time spectrum is of main importance in the perception of speech and that, the formants in the short-time spectrum are more important than other details of the spectral envelope. The first formant in the range of 250 to 800 Hz, is less important, than the second formant (Thomas, 1968). Thus it is possible to apply a certain degree of high pass filtering to speech which may affect the first formant without introducing serious degradation in intelligibility. Low-pass filtering with a cutoff frequency above 4 kHz will in general not seriously affect intelligibility. A good representation of the magnitude of the short-time spectrum is generally considered to be important whereas the phase is unimportant. Another aspect of the system that plays a role in speech enhancement is the ability to mask one signal with another. For example, narrow-band noise and many forms of artifical noise are more unpleasant to listen than broad-band noise and an enhancement system might include the introduction of broad-band noise to mask the narrow-band noise (Lim and Oppenheim, 1979).

The comparison concerned nonsense syllables, monosyllabic words, and words spoken both in isolated and in predictable and unpredictable sentence contexts (Dermody, 1992). When nonsense syllables are presented in continuous white noise, listeners recognize syllables at performance levels ranging from 45% to 80% over SNR from -5 dB to +5 dB. For monosyllable words presented in white noise the results range from 32% to 80%. Syllable recognition at 0 dB SNR is approximately equal to syllable recognition of machine speech recognizers in noise-free conditions. A communication involves both a speaker and a listener, speech communication in noise also has to deal with variables from the speaker, the listener, the task and the environment. In extremely noise environments it was found that replacing high energy consonants by white noise in a speech signal where high energy consonants have been filtered out could lead to almost full intelligibility for a certain noise level.

When speech is produced in noise, the type of masking noise, the sex of the speaker, and the type of vocabulary are also factors influencing the intelligibility. The multitalker babble noise degrades the intelligibility more than does white Gaussian noise for the digit vocabulary. The white Gaussian noise affects the consonants more than does multitalker bable noise, whereas multitalker noise affects the vowel portions of the words more than does white Gaussian noise. When speech is produced in noise female speakers seem to be more intelligible than male speakers. Conversation or interfering speech arises from other voices picked up by the microphone. Resolution of two or more voices in a conversation (cocktail effect) is a process inherent in the human speech understanding mechanism. Humans are able to concentrate on one voice to the exclusion of the others.

Noise adaptation is another phenomenon which might contribute to the ability of human listeners to understant speech in the presence of noise. It was shown that human listeners adapt to continuous noise (Ainsworth and Pratt, 1993). The same effect was replicated with speech recognition experiments involving an auditory model. Besides the factors already considered, a number of other operations (e.g., frequency and amplitude distortions) can modify the speech signal and thereby affect the intelligibility (Junqua and Haton, 1996).

2.4. *Performance Evaluation*

The performance evaluation of the various systems is a very difficult task, partly because the performance of a system may vary depending on the particular applications. Some systems which increase speech quality may decrease intelligibility. A further factor in evaluating the system performance is that the objective of various systems is an improvement in some aspects of human perception such as an improvement in intelligibility or quality, or reduction of listener fatique. Since the human perceptual domain is not well understood, a careful system evalution requires a speech intelligibility or quality tests. A careful subjective test can be tedious and time consuming, and requires processing a large amount of data.

In a *speech intelligibility test* (Hecker and Guttman, 1967), listeners are presented with test material and asked to identify the material or answer questions based on the test material. For example, listeners may be presented with sentences, words or syllables and asked to write the test material that they heard or choose one out of several options which most closely resembles what they heard. From the responses of the listeners the intelligibility score, the percentage of correct answers based on some predetermined criterion, is computed. For a given type of degradation, the intelligibility is obtained for several different levels of degradation.

The *amount of degradation* is represented in terms of SNR. For the same type and level of degradation, the intelligibility can vary depending on the test procedure, test material, training of persons, etc. Furthermore, the SNR employed varies from one evaluation to another. Two systems evaluated differently and with a different SNR cannot be compared. It is established that if one system is superior to another when evaluated by the same test, a similar result also holds when evaluated by different types of intelligibility tests (Lim and Oppenheim, 1979).

3. Intelligibility Enhancement System

The intelligibility enhancement system is based on spectral subtraction techniques (Ephraim and Malah, 1984; 1985). A noisy speech passes through the low-pass filter, which restricts the frequency band up to 4 kHz. Such frequency is selected because the limitation of the frequency band up to 4 kHz does not decrease the intelligibility. It suffices to sample the noisy speech with 8 kHz frequency. However to avoid the aliasing, the noisy speech is sampled with 10 kHz frequency.

One of the assumptions made is that the noise and the speech are uncorrelated and additive. The spectral subtraction techniques require estimation of the noise during pauses. It is supposed that the noise characteristics change slowly. This method necessitates the availability of an endpoint or a voice activation detector to separate speech from noise.

Noise spectral characteristics are estimated during silence intervals where only the noise is acting. The silence intervals determination algorithm establishes the borders between the neighbouring words. It is important for nonstationary noise to determine the threshold size. If the threshold is selected great, the noise is suppressed well, but the speech segments with low energy are distorted. Basically, the consonants are distorted on which the intelligibity strongly depends. If the threshold is small, the noise is suppressed little.

In the spectral analyzer of the noisy speech, the spectral density of the noisy speech is estimated. In the spectral analyzer of the noise, the spectral density of the noise is estimated. In the case where the noise is nonstationary, the spectral density is estimated in every silence interval, and after each estimation the coefficients of the filter are adjusted. The controller adjusts filter coefficients at every time instant. At the time instant, from the ith channel of the noise spectral analyzer we obtain the ith component of noise spectral density D_i, and from the ith channel of the noisy speech spectral analyzer we get the ith component of the noisy speech spectral density. Thus, we can calculate the spectral density of a clean speech signal in the ith channel at the kth time instant.

$$s_i^2 = y_i^2 - n_i^2.$$

If $y_i \leqslant n_i$, then

$$s_i = \begin{cases} \sqrt{y_i^2 - n_i^2}, & \text{if } y_i > n_i, \\ 0, & \text{if } y_i \leqslant n_i. \end{cases}$$

Then the adjustment coefficient at the kth time instant in the *ith* channel of filter is calculated by

$$\gamma_i = s_i^2/y_i^2 = (y_i^2 - n_i^2)/y_i^2.$$

The filter consists of 16 or 32 connected in parallel bandpass filters, which are uniformly steadily arranged in the desired frequency band from 100 up to 4000 Hz. Central frequencies and passbands of each bandpass filter are the same as the central frequencies and passbands of the spectral analyzers of noise and noisy speech. The input of the filter is a noisy signal and the output of the filter is a filtered noisy signal. At every time instant, the controller calculates the coefficients γ_i, which are used for determining the bandpass filter coefficients of amplification. The output of the filter is obtained by summing up all the bandpass filter outputs.

Passbands of the bandpass filters and their central frequencies are chosen so that the frequency response of the filter would be approximately equal to the mean spectral density of the speech in case the coefficients γ_i are not adjusted. When the coefficients γ_i are adjusted, the changes in the speech spectral density lead to the redistribution of the coefficients γ_i and to the modification of the filter frequency response. The frequency response of the filter will follow the changeability of the speech signal spectral density.The number of bandpass filters is defined with a desired accuracy (Kazlauskas *et al.*, 1985).

To implement the intelligibility enhancement system, a multitarget algorithm and a program for defining the digital filter coefficients have been developed and realized by computer. The program is devoted to the generation of spectral analyzer coefficients for the bandpass filters. The program is used for the calculation of the bandpass, low-pass, high-pass, and notch filters. Notch filters are used for filtering one or more undesirable sinusoidal frequencies. The program generates the coefficients for the cascade and parallel digital filters. The analysis of the spectral analyzer realized using the bandpass filters with 4 poles each showed that the definition of time-varying spectral density of the speech signal was of contradictory character. On the one hand, it is required that the passband of the bandpass filter be sufficiently narrow for ensuring reliable estimation of the speech signal spectral density in the narrow frequency band. On the other hand, to preserve the slowly time-varying components of the speech signal, the passband of the bandpass filters must be as wide as possible. The contradictions also exist in selecting the frequency band of the speech signal when the spectral density is estimated. With an increase of the frequency band the trend error also increases while the stochastic error reduces. Short-time spectral estimates obtained using the FFT algorithm are more accurate than the spectral estimates obtained using the bandpass filters. However, the spectral analyzer which is realized using the bandpass filters is more useful than that obtained by means of FFT in case, we need to compute a spectrum at every time instant. We use FFT in the intelligibility enhancement system. This choice is due to the fact that the short-time spectral estimation accuracy mainly determines the success of noisy speech filtering. The short-time spectral estimation by FFT is based on the analysis of the speech signal in the segmental manner. The speech signal is divided into the overlapping blocks of finite length. Afterwards, the segments are weighted and the spectrum of each segment is estimated (Kazlauskas *et al.*, 1988).

Another important problem is determination of the speech silence intervals. The result of noisy speech filtering strongly depends on the silence intervals problem solution. To determine the noisy speech silence intervals we used the following technique. We estimated the noisy speech spectral density at some time instant. Afterwards, we constantly processed the noisy speech and established spectral density changes. Assuming the spectral density as the main characteristic, we used the spectral error measure

$$SE = \left(\frac{1}{r(0)} - 1\right)^2 + 2\sum_{k=1}^{M} \frac{r^2(k)}{r^2(0)},$$

where $r(k)$ is the autocorrelation function of the speech signal; the first member on the right-hand side shows changes in the noisy speech spectral density while the second member shows the changes in the form of the noisy speech spectrum.

Vocal tract characteristics and intelligibility. In our experiments the problem was formulated so that we had not only the noisy speech signal but also the clean speech of the same person. So the question arose whether it was possible to use the vocal tract characteristics to increase speech intelligibility. For this purpose the computer program VINI was written. The program was based on the idea that using the clean speech the amplitude frequency characteristics of the vocal tract were computed. Afterwards, these characteristics were used in the noisy speech filtering process. Listening to the noisy speech filtering recorders showed that utilization of the vocal tract characteristics in the filtering process did not improve the intelligibility. This may be explained in such a way: the word sense lies in the short-time spectrum of the speech signal, not in the averaged noisy speech spectrum from which we calculated amplitude frequency characteristics and used in the filtering process.

3.1. *Estimation of Noise Spectral Density*

To estimate the spectral density of noise we use the method of periodograms (Bentkus, 1984). In the case of *stationary noise*, the spectral density is defined by

$$\overline{g}(\nu) = \frac{1}{\varkappa + 1} \sum_{l=0}^{\varkappa} J_n(\nu, l), \quad 0 \leqslant \nu \leqslant \frac{1}{2T},$$

where $1 \leqslant M \leqslant N,\ 1 < n < N,$

$$\varkappa = \begin{cases} (N-n)/m, & \text{if } (N-n)/m \text{ is integer,} \\ [(N-n)/m] + 1, & \text{in other cases.} \end{cases}$$

The normalized periodogram $J_n(\nu, l)$ is constructed using the segments $X_{lm+1}, \dots, X_{lm+n}$, i.e.,

$$J_n(\nu, l) = \frac{T}{NS_{x,l}^2}\left|\sum_{k=1}^{n} q_k(X_{lm+k} - \overline{X}_l)e^{-i2\pi kT\nu}\right|^2, \quad \nu \geqslant 0,$$

in which

$$\overline{X}_l = \frac{1}{n}\sum_{k=1}^{n} X_{lm+k}, \quad S_{x,l}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_{lm+k} - \overline{X}_l\right)^2,$$

N is the length of the noise realization, m is the segment shift size, ν is the frequency, T is the discretization period, q_k are the values of the window.

In the case of *nonstationary noise*, the spectral density is defined as

$$\overline{g}(\nu, l) = \frac{1}{\varkappa + 1}\sum_{i=0}^{\varkappa} J_n(\nu, l - i), \quad l = 0, 1, \dots,$$

where l is the spectral density computation time related with the real time t $(t = 0, 1, \dots)$ as $t = lm + n,\ n = 1, 2, \dots, N$;

$$J_n(\nu, l) = \frac{T}{NS_{x,l}^2}\left|\sum_{n=1}^{N} q_n\left(X_{lm+n} - \overline{X}_l\right)e^{-i2\pi nT\nu}\right|^2.$$

REMARK. It is recommended to compute the value $\varkappa$ from the relation $\varkappa + 1 = 2^M$, M is integer. We can compute the values $\overline{g}(\nu, l)$, if $\varkappa$, $\varkappa + 1, \dots$. The condition $J_n(\nu, l - i) = 0$ must be satisfied if $l - i < 0$.

3.2. *Spectral Density Estimation Device*

The spectral density estimation device is used to estimate the nonstationary noise in real time. In such a case, the buffer memory is used in the input of the device, in which the noise values are preserved. The noise discrete values calculator is connected with the input register. When 128 discrete values of noise are received in the input register, the calculator gives the control signal to begin computations according to the algorithm:

1. The mean value of the lth segment of noise is computed by

$$\overline{X}_l = \frac{1}{N}\sum_{n=1}^{N} X_{lm+n}, \quad N = 256,\ m = 128,\ l = 0, 1, \dots.$$

This operation is performed in the summator with shifting.

2. In each channel the differences $X_{lm+n} - \overline{X}_l$ $(n = 1, 2, \dots, 256)$ are computed.

3. The variance of the lth segment is computed by

$$S_{x,l}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_{lm+n} - \overline{X}_l)^2, \quad N = 256,\ l = 0, 1, \dots.$$

This operation is performed in the summator with shifting.

4. The normalization coefficient $T/NS_{x,l}^2$ is computed, where T is the discretization period $(T = 1)$, N is the segment size $(N = 256)$.

5. The differences $X_{lm+n} - \overline{X}_l$ in all 256 channels are multiplied by the window values $q_1, \ldots, q_N$.

6. The values $q_n\ (X_{lm+n} - \overline{X}_l)$ are used as input values of the 256 point FFT processor.

7. In each output of the 256 point FFT processor, we obtain real Re$X_{l,n}$ and imaginary Im$X_{l,n}$ values, which are used for the calculation of Re$^2X_{l,n}$+Im$^2X_{l,n}$ in each channel of the device.

8. The multipliers are used in each channel and the values Re$^2X_{l,n}$+Im$^2X_{l,n}$ $(n = 1, \ldots, 256)$ are multiplied by the normalization coefficient $T/NS_{x,l}^2$. In the outputs of multipliers we get the values of the normalized periodogram.

9. The values of the normalized periodogram $J_n(\nu, l)$ at the lth time moment act on the N summators, and spectral density estimates are computed according to the formula:

$$\overline{g}(\nu, l) = \frac{1}{\varkappa + 1} \sum_{i=0}^{\varkappa} J_n(\nu, l - i).$$

If $\varkappa + 1 = 2^M$, M being integer, then division is changed by a shifting operation. The summators in each channel must have memory to store $\varkappa$ values of the normalized periodogram.

The modelling results show that in order to efficiently estimate the noise spectral density, it suffices to use 8192 values of noise (length of a realization is about 0.5 sec). Then, with the segment size $N = 256$ and segment shifting size $m = 128$, we obtain that $\varkappa + 1 = 8192/128 = 64$, and $\varkappa + 1 = 64 = 2^6$, $M = 6$.

The spectral density estimation device without any essential changes may be used to estimate the short-time spectral density of *noisy speech*. It is known that a nonstationary speech signal in short-time intervals of the 20 msec. length may be in fact considered as a stationary signal, i.e., in the interval of this length can use smoothing. The speech signal was discretized by 15000 Hz frequency, i.e., during 20 msec. 300 noisy speech values receive the input of the spectral estimation device. In order not to change the device, we consider that the block length is $N = 256$. If smoothing is performed during the 20 msec. length intervals, then $\varkappa + 1 = 2^M$, $M = 1$, $\varkappa = 1$ and the device will perform smoothing using two segments. If we assume the speech signal to be stationary in the 40 msec. intervals, then we can perform smoothing of 512 values, thus $\varkappa = 3$, and $M = 2$. In such a case, we perform smoothing of 4 segments (segment size is $N = 256$) with a segment shift $m = 128$. In order to get more efficient estimates of the speech spectral density, it is necessary to use a summator and perform additional smoothing with the Hanning window in each channel of the spectral density estimation device.The power of the nonstationary noise power spectral density depends on frequency and time, consequently, the quality of estimation of power of the spectral density is mainly defined by the convergence rate of estimations for each frequency. For the white noise, it is established that the convergence rate of the estimates of smoothed periodograms is higher than that of the coloured noise. For the white noise, the convergence of the power spectral density for all frequencies is approximately the same and already after 0.25 sec. (4096 noise values) the estimates are good.

The spectral density estimation device is based on the algorithm of smoothed periodograms. The algorithm differs from that of adaptive histograms in which all the periodograms that we get in the processing process are smoothed in each channel separately. In the algorithm of adaptive histograms, the histograms are estimated using the periodograms of each channel and according to the histograms the solution is determined on the power noise spectral density in each channel. The convergence rates of the noise power spectral density are different in separate channels of the device. The modelling results show that in some channels, the estimates converge to true values in 0.1 sec., while in other channels the convergence time of estimates is about 1 sec. On the average, in all the channels the convergence time of estimates is less than 0.5 sec.

In solving the problem of noise spectral density estimation it is also necessary to give the convergence estimation accuracy, because the estimate convergence time depends on the desired estimation accuracy. The variance is reduced about 2 times if the segment overlapping is 1/2 as compared to the case without segment overlapping. The spectral density estimation device is adaptive, in the sense that the input register continuously receives noise values from the silence intervals determination block and the spectral density estimates are computed approximately in 0.01 sec.

From the standpoint of realization, the estimation algorithm is simpler as compared to the adaptive histograms algorithm, because it is easier to compute the mean rather than the histograms in each channel and to use them for computing the spectral density.To compute the mean in each channel is rather simple, if in all the channels we compute values of the periodogram, where $\varkappa = 2^M$ (M is integer). In such a case, the mean and hence the power spectral density estimate in each channel is computed by accumulation of $\varkappa$ values of the periodogram with M shifting in the register.

The modeling results of the noise power spectral density algorithm show that it is possible to design a device using the chips of series K1815 and that good estimates will be obtained in the noise intervals of about 0.5 sec., i.e., the device can estimate the noise spectral density, noise nonstationarity being approximately 1 sec. The smoothed periodogram method has a good theoretical foundation, therefore it is more reliable as compared to the adaptive histograms method. The smoothed periodogram algorithm may be applied to any noise as well as in the estimation of the noisy speech short-time spectrum (Kazlauskas *et al.*, 1988).

3.3. *Determination of Silence Intervals*

In the problem of determination of silence intervals the main tasks are: 1) to choose a criterion by which the silence intervals are determined; 2) to choose the threshold.

The following characteristics may be used:

1. Energy of a signal in the interval $E = \sum\limits_{i=1}^{N} s^2(i)$;

2. Logarithm of the energy $E_l = \log E$;
3. Variance in the interval $\sigma^2 = \frac{1}{N} \sum\limits_{i=1}^{N} s^2(i)$;
4. Logarithm of the variance $P_\sigma = \log \sigma^2$;
5. The number of passages of the signal via the axis x in some interval;
6. The normalized correlation coefficient

$$C_1 = \frac{\sum\limits_{i=1}^{N} s(i)s(i-1)}{\sqrt{\sum\limits_{i=1}^{N} s^2(i) \sum\limits_{i=1}^{N-1} s^2(i)}};$$

7. The sum of correlation function samples $C_2 = \sum\limits_{i=1}^{m} r^2(i)$.

The first four characteristics are closely related and reflect the signal energy in the interval. It is known that the energy of vowels is noticeably higher than the energy of consonants. The 5th characteristic indicates the frequency at which the energy of the signal is concentrated. The 7th characteristic determines the behaviour of the correlation function.

The most perspective characteristics for our problem are the energy and the logarithm of energy, and the sum of correlation function samples. Experimental investigations show that the 5th and 6th characteristics are useless. The 7th characteristic may be used only in the case, where noise is uncorrelated. If the noise is correlated, then the noise and speech signal correlations are similar.

Modelling shows that when SNR= $0 \div -5$ dB, the energy can be used as characteristic for establishing the silence intervals. A more suitable characteristic is the energy logarithm. After the energy in the interval is determined, it is necessary to perform additional smoothing.

The choice of the threshold is a complex task even in the case where processes are stationary. The speech signal is a nonstationary process, so finding of the threshold is more complicated. The difficulty of threshold finding consists in the fact that the threshold size depends on the statistical properties of noise. However, the properties of noise are not known. Therefore, we have no other choice than to experimentally define the threshold size in such a way: to add the noise with known statistical characteristics to a clean speech signal, and then define the threshold size. To repeat this procedure for N times with different noise statistical characteristics and to define threshold sizes. Using these experimental data we can obtain the relationship among the threshold size and the noise statistical characteristics.

The modelling results show that it suffices to use a segment 20 msec. length (i.e., 300 values of noisy speech). The segments can be used without overlapping. To determine the silence intervals, we recommend using the energy or the energy logarithm to find the threshold. To determine the silence intervals the normalized correlation coefficient can be used only if the noise is uncorrelated. In such a case, it is necessary to keep in mind that the noise and noisy speech correlation function values are not distinguished in the intervals with sibilants. After the noisy energy as a function of the interval number is determined, the low-pass filtering is necessary. It is recommended to choose the threshold size so that the noisy speech periodogram be divided into 80% and 20% parts (Kazlauskas *et al.*, 1987).

3.4. *Efficiency of Silence Interval Determination Algorithms*

It is impossible to define the efficiency of silence interval determination algorithms analytically, because it is necessary to have an exact mathematical description of a nonstationary speech signal. Thus, it remains only the experimental way of estimating the efficiency of silence interval determination algorithms. The efficiency of these algorithms is defined as follows:

1. For the clean speech signal, we estimate the beginning and the end of each word.
2. The clean speech signal is added with noise, statistical characteristics of which are known.
3. We use the silence interval determination algorithm and estimate the beginning and the end of each word.
4. As a measure of efficiency of the algorithm the expression:

$$Q_\sigma = \frac{1}{N}\sum_{i=1}^{N}|H_i - H_i^*| + \frac{1}{N}\sum_{i=1}^{N}|K_i - K_i^*|,$$

   is used, where H_i is the beginning of the ith word, K_i is the end of the ith word, H_i^* is the beginning of the ith word, which is defined using the silence interval determination algorithm, K_i^* is the end of the ith word, defined using the silence interval determination algorithm.
5. It is possible not to restrict oneself with that and to continue the experiment of efficiency estimation of the silence interval determination algorithm, i.e., to repeat items 3 and 4 M times with different statistical properties of noise. In such a case, the measure of algorithm efficiency is:

$$Q = \frac{1}{M}\sum_{\sigma=1}^{M} Q_\sigma,$$

   which defines the efficiency of the silence interval determination algorithm and does not depend on the statistical properties of noise.

Another approach to the estimation of efficiency of the silence interval determination algorithm is used for the algorithms that are based on the classification methods:

1. For the clean speech signal, the number of words and silence intervals is calculated.
2. The clean speech signal is added with noise statistical characteristics of which are known.

3. The silence interval determination algorithm is used and the number of mistakes in the classification of words and silence intervals is calculated. The measure of the algorithm efficiency is the number of mistakes.
4. For the measure to be independent of the noise statistical properties, it is necessary to repeat item 3 M times with different statistical properties of noise. Then, the measure of algorithm efficiency is

$$\theta = \frac{1}{M}\sum_{\sigma=1}^{M}\theta_{\sigma},$$

where θ_{σ} is the number of mistakes for noise with the given statistical properties.

By comparing measures Q and θ, we estimate which silence interval determination algorithm is better.

3.5. *Signal to Noise Ratio Evaluation*

Signal to the noise ratio (SNR) is defined as follows

$$SNR = 10\log(\sigma_s^2/\sigma_n^2),$$

where σ_s^2 is signal variance, σ_n^2 is noise variance. The noise variance is estimated as an average over a segment of the signal which contains only noise. The measure depends on how much silence the speech signal contains, and it also assumes that the speech and noise signals are *stationary*. For *nonstationary* speech, the SNR measure takes into account the fact that the same amount of noise has different values depending on the signal level. The segmental SNR is defined as in (Junqua and Haton, 1996).

$$SNR = \frac{10}{M}\sum_{m=0}^{M-1}\log\left(\sigma_{s,m}^2(N)/\sigma_{n,m}^2(N)\right),$$

where N is the segment length, M is the number of segments, $\sigma_{s,m}^2(N)$ is the speech signal variance, $\sigma_{n,m}^2(N)$ is the noise variance.

The segmental SNR assigns equal weights to both loud and soft portions of utterance. It is based on a log weighting which converts SNR values to dB values prior to averaging, so that very high SNR segments are not over emphasized as compared to low SNR segments. Another measure is the maximum SNR in dB. The advantage of such a measure is its independence of the amount of silence found in the speech signal.

The disadvantage of all these measures is that none of them provides information about the shape of the spectrum and how speech and noise are distributed across frequency bands.

SNR definition. In modelling problems it is necessary to define a desired SNR in dB. For this reason a coefficient k is used. We define the relationship of the coefficient k with the SNR in dB. Then the SNR in dB is defined by

$$SNR = 10\log\left(\sigma_s^2/k^2\sigma_n^2\right)$$

or

$$SNR/20 = \log\left(\sigma_s/k\sigma_n\right),$$

and

$$k = (\sigma_s/\sigma_n)\cdot 10^{-SNR/20}.$$

Thus, we can compute such a coefficient k that is multiplied by each noise sample to get the desired SNR in dB.

4. Speech Intelligibity Enhancement Device

To make experiments on the speech intelligibility enhancement, an engineer V. Steponėnas have designed and developed the device (Kazlauskas *et al.*, 1985). Fig. 1 shows the structure of the speech intelligibility enhancement device. The device has two equal channels. Each channel consists of a low-pass and a high-pass filters. By means of these filters we can filter a low-frequencies up to 25, 98, 131, 196, 262, 392, 524, 784, and 1046 Hz, and a high-frequencies up to 1046, 1568, 2093, 3136, 4186, 6272, 8372, 12544, and 16744 Hz. Active filters sharpnesses in the transition passband is 12 dB/octave. The device can raise high-frequencies of the speech signal in the passband from 131 Hz up to 4186 Hz with 2, 4, 6, 8 dB/octave.

For preliminary amplification of the signal, the two-cascade amplifiers with $SNR =$ 78 dB are used. The transfer coefficient is 40 dB. The spectral passband is from 25 Hz up to 25 kHz. High frequencies are amplified in differential blocks. An increase in the amplification is 2, 4, 6, and 8 dB/octave, starting from 262 Hz up to 4192 Hz. Transfer coefficients of the low-pass and high-pass filters are from 1.0 up to 1.3, and the sharpness is from 12 up to 15 dB/octave. For low-pass and high-pass filters 10 fixed cutoff frequencies are chosen. So we can produce 100 filters. Stepped variation of the speech signal level with 5dB jumps from 5 dB up to 55 dB is performed using the attenuator. It is possible to add the clean speech signal with the white noise. For that purpose the white noise generator is used.

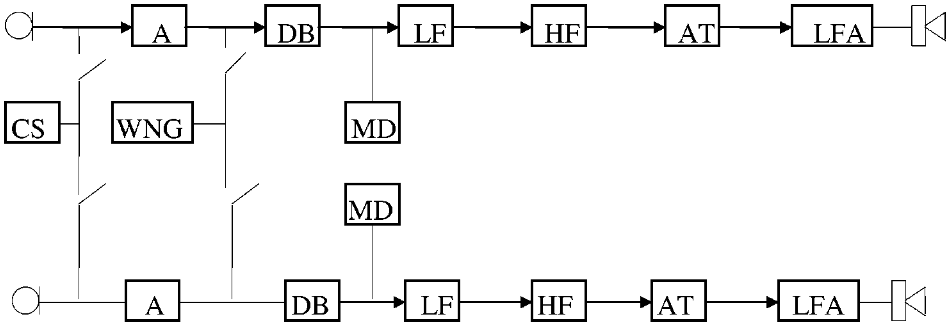

Fig. 1. Structure of the speech intelligibility enhancement device. A – amplifier, LF – low-pass filter, AT – attenuator, MD – measure device, CS – control signal, DB – differential block, HF – high-pass filter, LFA – low-frequency amplifier, WNG – white noise generator.

Experimental results. The purpose of the experiment is to explore how the cutoff of low frequencies and high frequencies, and amplification of high frequencies of the noisy speech changes the intelligibility. The experiment was carried out in the 30 m^2 room. In the first four cases, the announcer read aloud some words from special tables. The microphone was installed in a fluorescent lamp, and the announcer was 1 m away. In the 5th case, we filtered sentences. By varying the cutoff of low- and high-frequencies as well as the amplification of high frequencies we investigated the influence on the noisy speech intelligibility. It was found that the noisy speech intelligibility increased from 5% up to 8% depending on the circumstances. Hence, we did not obtain a considerable visible speech intelligibility enhancement. However, using this device we can quickly change the frequency band and amplification, and repeatedly listen to the same text in different frequency bands. That helps a listener to better understand the noisy speech. In the fluorescent lamp noise case, it is not sufficient to filter low and high frequencies, because the intelligibility increases imperceptibly. In such a case, we additionally used notch filters, and the intelligibility increased up to 14%. In the 6th experiment, we investigated a speech signal of two or three simultaneously speaking announcers. The speech intelligibility increased after we have filtered low frequencies up to 392 Hz. The intelligibility of a woman's voice increased after amplifying the high frequencies (increase in high frequencies was 4 dB/octave), and the frequency band was restricted up to 392 Hz. It was estabilished that the intelligibility increased in the case where the frequencies higher than 2093 Hz were eliminated (Kazlauskas *et al.*, 1985).

5. Conclusions

An important factor for increasing intelligibility is the restoration accuracy of time and frequency envelops of the noisy speech. Under the noise influence the envelopes of the sounds in time and frequency domains are changed. In the first place, the fluctuating noise distorts voiceless consonants, afterwards, the sibilants, voiced consonants, and vowels. The speech signal is distorted most by the speech noise. Noises that act in the frequency band of the speech have the main influence on the intelligibility. In the case of additive noise, the frequency band restriction from 300 Hz up to 3500 Hz increases the intelligibility. Further narrowing of the frequency band reduces the intelligibility. In the noisy speech amplitude restriction, the pulsation effect of the noise and speech is observed. This effect is caused by mutual modulation of the noise and speech, and reduces the speech intelligibility. The narrow band noise is more unpleasant for hearing than the wide band noise.

A speech signal is modelled as a nonstationary normal stochastic process with a slowly varying variance and spectral density. The speech signal correlation interval is 2–3 msec. The length of sounds is from 25 up to 250 msec. The average length of the sound is 125 msec. The man pronounces sounds and words without pauses, which occur only for breath between the word groups or phrases. The consonants take a relatively little part of the speech signal energy, but they are important for intelligibility. The sounds

p, n, m, t, v, c, i bear more than 50% semantic information. For vowels the hearing does not perceive distortions of the spectral envelope in considerable limits. For the consonants intelligibility, the time distortion of the formant maxima and spectral energy changes in the narrow bands are of importance. The short-time spectrum of the speech is of most significance. The first formant is less important than the second one. The phase spectrum has little significance. The pauses among the words are used for the definition of noise spectral characteristics. When a man is speaking, approximately 30% of time are pauses. In the case of continuous speech, the word segmentation is a difficult and often nonsolvable problem.

The intelligibility enhancement filters must be based on the speech perception features. Low-pass filtering with the cutting frequency up to 4 kHz affects the speech quality and the intelligibility essentially does not become worse.

If the low-pass filter is used for several simultaneously talking announcers, we observed an increase in the intelligibility when the frequencies from 2100Hz and higher were eliminated. Application of the announcer's voice mean spectral characteristics to the filtering of noisy speech does not increase the intelligibility.

In the case where the speech signal is corrupted by the noise of the flourescent lamp, then for increasing the speech intelligibility it does not suffice to filter and amplify low and high frequencies.

The proposed digital filter design techniques is simple and easily realizable. A program generator according to the desired frequency characteristics calculates the coefficients of the low-pass, high-pass, and bandpass filters. For the bandpass filter with small passbands, the poles are near the nonstability bound.

Determination of the time-varying spectral density is related with overcoming of a contradiction in selecting a passband of the bandpass filter. On the one hand, it is demanded that the passband would be sufficiently narrow to ensure a good estimate of the speech spectral energy in the narrow passband. On the other hand, to preserve components of the noisy speech time trend, the passband must be as wide as possible. The spectral analyzer with bandpass filters is more useful than FFT in the case where it is necessary to calculate the time-varying spectrum at every time instant.The spectral analyzer with bandpass filters demands less computations. However, the short-time spectral analysis by FFT yields more exact estimates of the energy spectrum than the spectral analyzer with bandpass filters.

The analysis of different filtering methods of noisy speech shows that a priori information on SNR is necessary. Such uncertainty is defined a priori by an unknown parameter, that is chosen as a constant one or is changed in the task solving time. Many filters increase SNR that increases the speech quality, but does not increase the intelligibility. Many considered methods are based on the mathematically optimal methods, such as the mean square error minimization or the probability distribution function maximization. However, these criteria do not reflect the peculiarities of speech perception. It is necessary to prepare such a mathematical criterion of error, that would estimate the speech audibility features. The problem of enhancement of noisy speech intelligibility is important and needs new approaches and principles.

6. Acknowledgments

I would like to thank Associate Professor C. Paulauskas, engineer V. Steponėnas, and programmer G. Dagytė for their active participation in performing this work.

References

Abel, S. M., and P. W. Alberti (1982). Speech intelligibility in noise: Effects of fluency and hearing protector type. *J. Acoust. Soc. Am.* **71**(3), 708–715.

Ainsworth, W., and S. Pratt (1993). Comparing error correction strategies in speech recognition systems. In C. Baber and J. Noyes (Eds.), *Interactive Speech Technology*. pp. 131–135.

Bentkus, R. (1984). Optimal statistical estimates of the spectral density in L_2. *Lithuanian Mathematical Journal*, **24**(3), 51–69.

Curtis, R.A., and V. Neiderjohn (1978). An investigation of several frequency domain processing methods for enhancing the intelligibility of speech in wideband random noise. In *Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing*. pp. 606–609.

Dermody, P. (1992). Human capabilities for speech processing in noise. In *ETRW: Speech Processing in Adverse Conditions*. pp. 11–19.

Ephraim, Y., and D. Malah (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech and Signal Processing*, **32**(6), 1109–1121.

Ephraim, Y., and D. Malah (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech and Signal Processing*, **32**(2), 443–445.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Lexington, M.A.

Flanagan, J.L. (1972). *Speech Analysis, Synthesis and Perception*. New York: Springer-Verlag.

Hecker, M.H., and N. Guttman (1967). Asurvey of methods for measuring speech quality. *J. Audio Eng. Soc.*, **15**(5), 400–413.

Jungua, J.C., and J.P. Haton (1996). *Robustness in Automatic Speech Recognition*. Kluwer Academic Press.

Kazlauskas, K., C. Paulauskas, V. Steponėnas and G. Dagytė (1981–1988). Investigation of the speech intelligibility enhancement methods. *Research Reports*. Vilnius.

Lim, J.S., and A.V. Oppenheim (1979). Enhancement and bandwidth compression of noisy speech. *IEEE*, **67**(2), 1586–1604.

Rabiner, L.R., and R.W. Schafer (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall.

Sambur, M.R. (1978). Adaptive noise cancelling for speech signals. *IEEE Trans. Acoust., Speech and Signal Processing*, **26**(6), 419–423.

Thomas, I.B. (1968). The influence of first and second formants on the intelligibility of clipped speech. *J. Audio Eng. Soc.*, **16**(2), 182–185.

Thomas, I.B., and A. Ravindran (1974). Intelligibility enhancement of already noisy speech signals. *J. Audio Eng. Soc.*, **22**(3), 234–236.

K. Kazlauskas received Ph.D. degree from Kaunas Polytechnic Institute (Kaunas, Lithuania) in 1975. He is a senior researcher of the Process Recognition Department at the Institute of Mathematics and Informatics and Associate Professor at the Vilnius Pedagogical University. His research interests include design of concurrent algorithm and architectures for signal processing, and computer aided design of signal processing systems.

Užtriukšmintos kalbos suprantamumo pagerinimas

Kazys KAZLAUSKAS

Straipsnis skirtas užtriukšmintos kalbos signalo suprantamumo pagerinimo tyrimui. Trumpai apžvelgiama kalbos signalo modelis, triukšmo šaltiniai, suprantamumas ir jo kokybės įvertinimas. Analizuojama kalbos suprantamumo pagerinimo sistema, besiremianti spektrų skirtumų metodu, ir spektrinio tankio įvertinimo įrenginys, kuriame naudojamas suglodintų periodogramų algoritmas. Aptariama signalo ir triukšmo santykio įvertinimas bei intervalų tarp žodžių nustatymas ir jo efektyvumas. Aprašomas kalbos suprantamumo pagerinimo įrenginys.