THE NORMALIZATION OF THE SPEECH SIGNAL SPECTRUM ENVELOPE

ANDRZEJ PLUCIŃSKI

Institute of Linguistics
Adam Mickiewicz University
61-874 Poznań, Poland

ABSTRACT

This paper proposes normalization of the speech signal envelope by means of multiplicative centralization. The method proposed is based on the assumption that speech signal analysers of the human nervous system identify such instantenous speech signal spectra which can be superimposed by means of multiplicative transformations. The method doesn't involve any initial classifications (e.g. into male/female voices, vocalic/consonantal sounds, etc.). An alternative representation of the speech signal spectrum by means of coefficients of an extension of the speech signals spectrum envelope into a power series is suggested. Such a representation give us a possibility to get rid off stationary contributions.

INTRODUCTION

An analysis of the influence of various distortions on speech intelligibility may help us to discover some mechanismus of sound perception. Simple and frequently met distortions are introduced by e.g. our electroacoustic equipment. We have no difficulty in finding out that the level of recordings being reproduced has almost no influence on the reception of the content being transmitted. Also dislocations of the spectrum in the frequency range, due to the change of the speed reproduction can reach considerable values with no effect on intelligibility. As seen from a formal point of view, these distortions consist in a multiplication of amplitudes or frequencies of the spectral components by certain constants. Apart from the above mensioned distortions we can also find linear distortions which consist in the attenuation of various spectral components, especially the extreme ones. On the basis of such observations we can infer that the received accoustic signal undergoes certain normalization in the process of perception. The normalization allows us to compensate for the interpersonal differences and for the influence of the conditions of the acoustic wave propagation.

1. A MATHEMATICAL MODEL

The distortions of the signal caused by both the change of amplification level and a non-uniform transmission of spectral components (tone quality) pertain to amplitude and can be described by means of a function dependent on frequency. Amplitudes of signal spectral components will be multiplied by values of that function. A constant component of the function will be responsible for the general amplification level. Distortions in time can be shown as multiplication of the frequency scale by the constant. Let $a(f,t)$ be a dependence which shows an envelope of the spectrum amplitude of speech signal. In accordance with the above remarks we can say that

$$a(f,t) = b(f)\varphi(f\nu,t), \qquad (1)$$

where b is a dependence which describes the signal transmitling characteristic, φ is an envelope of the primary signal spectrum, t denotes time, f-frequency and ν is a constant responsible for the displacement of the signal spectrum in the frequency range. The concept of primary signal will be clarified in the subsequent part of the paper.

As a result of such a formulation of the mathematical model we will be claiming that the distortions in question consist in multiplicative transformations of the signal in the amplitude and frequency range. We will show now that, by using multiplicative centralization, Fourier transform of the signal can be reduced to a certain standard form, free from the influence of these distortions. The centralization consists in dividing the amplitude and multiplying the frequency of spectral components by an appropriate weighted arithmetic mean.

2. NORMALIZATION BY MEANS OF MULTIPLICATIVE CENTRALIZATION

Distortions in amplitude will be eliminated according to formula (1) by dividing the envelope of the Fourier transform of the signal by the arithmetic mean

$$a^{o}(f,t) = a(f,t)/\mu(f), \qquad (1)$$

where $\mu(f)$ designates the weighted arithmetic mean.

Distortions in frequency will be eliminated if we multiply the arithmetic mean of the result of the previous operation calculated in the frequency dimension. Thus we have

$$a^{N}(f,t)=a^{o}(f\mu(t),t).$$

To justify operation (1), let us calculate, without going into details, the time mean of the envelope of the speech signal spectrum:

$$\mu(f)=(\int_{t_d}^{t_g} w(t)dt)^{-1}\int_{t_d}^{t_g} w(t)b(f)\varphi(fv,t)dt.$$

Let us substitute W for $\int_{t_d}^{t_g} w(t)dt.$

Using the properties of the integral, we can write that

$$\mu(f)=W^{-1}b(f)\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt = C_{\varphi}(f)b(f).$$

Thus, we can notice that as a result of operation (1) the multipier $b(f)$ is removed. Let us now calculate the same mean for the centralized process:

$$\mu'(f)=W^{-1}\int_{t_d}^{t_g} w(t)a(f,t)/\mu(f)dt =$$

$$= W^{-1}\int_{t_d}^{t_g} w(t)b(f)\varphi(fv,t)/$$

$$(W^{-1}b(f)\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt)dt =$$

$$= (\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt)^{-1}\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt=1.$$

Hence, the next multiplicative centralizations will have no influence on the results. It follows from the above that the primary process φ is one which is invariable in relation to the multiplicative centralization of its amplitude, and that

$$C_{\varphi}(f) = 1$$

Let us calculate now the mean normalizing the position of the spectrum in the frequency dimension. Frequency is an independent variable; the only information on what frequency range the instantenous spectrum comprises is given to us by amplitudes of the components. For that reason it was decided to use them as a weight for each point in the frequency dimesnion. Such a selection of the weight function makes it possible to average only the process normalized in the amplitude dimension, because linear distortions will have a significant influence on the normalization in the frequency dimension. Thus we calculate

$$\mu(t)=(\int_{f_d}^{f_g}\varphi(f,t)df)^{-1}\int_{f_d}^{f_g}\varphi(f,t)fdf. \qquad (2)$$

If we further substitute f for s/v, we obtain

$$\mu(t)=(\int_{s_d}^{s_g}\varphi(s/v,t)/v ds)^{-1}\int_{s_d}^{s_g}\varphi(s/v,t)s/v^2 ds.$$

Using W for $\int_{s_d}^{s_g}\varphi(s/v,t)ds$, we write then that

$$\mu(t)= \frac{1}{v} W^{-1}\int_{s_d}^{s_g}\varphi(s/v,t)sds.$$

This result can be briefly written in the form

$$\mu(t) = \frac{1}{v} C_{\varphi_t}$$

If we then calculate the weighted mean (2) for the process already centralized in the frequency and amplitude domains, we will find that it will be equal to 1. It follows from it that

$$C_{\varphi_t} = 1 .$$

Thus, we can say that primary envelope φ of the speech signal spectrum is one which does not change under the influence of multiplicative centralization in the amplitude and frequency ranges.

3. SOME DETAILS OF THE NORMALIZATION PROCEDURES

I want to show now how to compute the means $\mu(f)$ and $\mu(t)$. It turns out that the computations will not be complicated when we use the extension to the power series of the spectrum envelope

$$a(f,t) = \sum_{i=0}^{I}\sum_{j=0}^{J}\alpha_{ij}f^{i}t^{j}.$$

3.1. Normalization in the amplitude range

Without affecting severely the previous considerations we can assume a week dependence of the function b on time. In consequence, a running mean $\mu(f,t_o)$ will be computed. t_o denotes the current moment.

The averaging of a signal by means of the running mean is equivalent to its filtering by means of a low-pass filter (Steiglitz 1977). In order to define the required averaging time, it suffices to determine the parameters of an equivalent filter. The parameters of an equivalent filter depend solely on the shape of the weighting curve w(t), i.e. on the so called time-window. Choosing the time-window of $\cos^2$ type we obtain the amplitude characteristic of the equivalent filter showed in fig. 1 (Pluciński 1986).

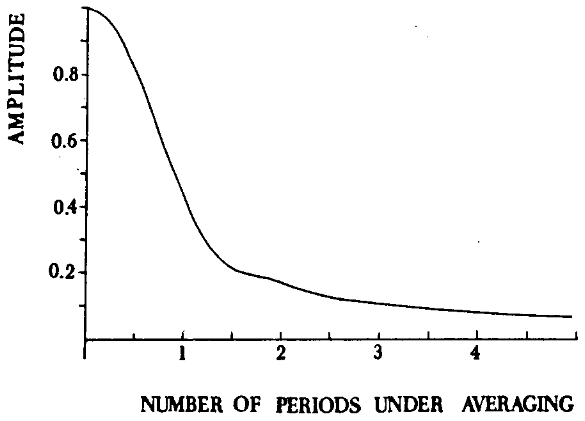

Fig. 1.

In effect the normalization, frequencies lying in the pass band of the equivalent filter will be compensated. Since the articulation time of individual speech sounds at normal speech tempo rarely goes over 100 ms, we can assume that changes in the spectrum envelope slower then 10Hz should be eliminated. Thus, as can be seen from fig. 1, the averaging time should not be longer than 200ms. The running mean for the time-window of $\cos^2$ type, i.e. if

$$w(t) = \cos^2(\frac{t-t_o}{t_o-t_d}\ \frac{\pi}{2}),$$

can be calculated according to the formula

$$\mu(f,t_o=0)= \sum_{j=0}^{J}((\frac{1}{j+1}(\frac{-1}{\pi})^{j+1}(j!\sin(\frac{\pi}{2}j) +$$

$$+ \sum_{k=0}^{j} k!\binom{j}{k}(-\pi)^{j-k}\sin(\frac{\pi}{2}k)))t_d^{j}\sum_{i=0}^{I}\alpha_{ij}f^{i})$$

(c.v. Ryżyk 1964: 132, formula 2.513.4).

3.2. Normalization in the frequency range

The weighted mean over the frequency can be calculated according to the formula

$$\mu(t_j)=(\sum_{i=0}^{I}\alpha_{ij}(f_g^{i+1}-f_d^{i+1}))^{-1}.$$

$$\cdot\sum_{i=0}^{I}\frac{\alpha_{ij}}{i+2}(f_g^{i+2} - f_d^{i+2}),$$

where f_g and f_d denote borders of integration in the frequency range. This range should cover the acoustic band.

4. A PARAMETRIC REPRESENTATION OF THE SPECTRUM

In the computational technique applied here a development of the envelope of the instantaneous spectrum into a power series is used. The coefficients of this development can be used for a parametrical representation of changes of the signal spectrum in time. This has the following adwantages:
1) it allows for an uniform representation of both vowels and consonants by means of a trajectory in the space of those coefficients,
2) all information on the spectrum envelope, i.e. on the position of both maxima and minima, their amplitudes and widths is contained in this representation.

One can also expect that coefficients α_{0j} and α_{1j} of this development will not have any significant influence on automatic speech recognition. Using such a representation, we can easily remove from the spectrum a time independent component represented by coefficients α_{i0}. In the representation proposed herein we simply reject these coefficients which results in elimination of stationary noises.

We can find the parameters of the running polynomial approximation calculated on the basis of n samples taken at equal (time) intervals by the analysis of characteristics of an equivalent digital filter (Pluciński (1986)). In fig. 2 there are shown amplitude characteristics of filters equivalent to running approximation by means of the third-degree polynomial (J=3) on the basis of seven samples. There are shown amplitude characteristics for three types of time-windows, namely for:
1) rectangular time-window, i.e. $w(t_k)=1$,
2) $\cos^2$ time-window, i.e.

$$w(t_k) = \cos^2((k-n)\pi/(2(n-1)),$$

3) Gauss time-window, i.e.

$$w(t_k) = \exp(((k-n)^2/(n-1)^2)\ln p),$$

where $k=1,\dots,n$, $n=7$, p stands for the cut-off level of the Gauss curve, $p=0.01$.

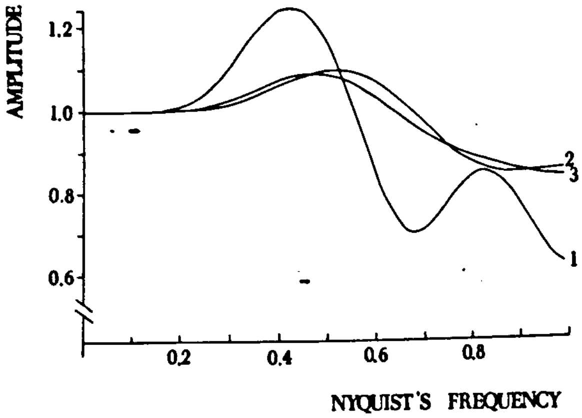

Fig. 2.

DISCUSSION

Hitherto known procedures for normalization of spectrum parameters concern formant frequencies. In order to motivate their proposals some authors refer to the anathomy of the organs of speech (Wakita (1977)), some authors to the properties of hearing (Syrdal (1986)) while others inform us only about the efficiency of a procedure of some kind (Lobanov (1971), Miller et al. (1980)). While forming normalizing rules, we aim at giving such rules which can help us identify some of numerical sets or sequences. Therefore, when analysing the rule that has been proposed by Lobanov, i.e. $F_i^N = (F_i - \overline{F}_i)/\sigma_i$, where $\overline{F}_i$ stands for a mean of i-th formant frequencies over all vowels and σ_i stands for standard deviation, we find out that this rule identify all numerical sets with elements y=ax+b, where a and b are any arbitrary constants. It may be proved by substituting F_i by aF_i+b. Analogously we can prove that Miller's formula - $F_{ij}^N = F_{ij} \prod_{i=1}^{3} F_{ij}^{-1/3}$, where i stands for the number of the vowel and j for the number of an observation - identify numerical sets with elements y=ax. The same concerns formulas proposed by (among others) Syrdal, Wakita and van Dijk. Like in the case of Miller's rule the procedure proposed in this paper identifies numerical sequences with elements y=ax. It is also a nonuniform procedure.

REFERENCES

Dijk, J.S. van. 1984. Conservation of vowel contrast in various speech conditions. Proceedings from the Institute of Phonetic Sciences of the University of Amsterdam, 8, 19-31.

Lobanov, B.M. 1971. Classification of russians vowels spoken by different speakers. J. Acoust. Soc. Am., 49, 606-608.

Miller, J.D., Engebretson, A.M., Vemula, N.R. 1980. Vowel normalization: Differences between vowels spoken by children, women and men. J. Acoust. Soc. Am., Suppl. 1, 68, S33.

Pluciński, A. 1985. The normalization of the speech signal spectrum; paper submitted to Studia Phonetica Posnaniensia 1, 57-68.

Pluciński, A. 1986. The parameters of running approximation and averaging; paper submitted to Studia Phonetica Posnaniensia, 2.

Ryżyk, I., Gradstein, I. 1964. Tablice całek, sum, szeregów i iloczynów. PWN. Warszawa.

Steiglitz, K. 1977. Wstęp do systemów dyskretnych. WNT. Warszawa.

Syrdal, A.K., Gopal, H.S. 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. J. Acoust. Soc. Am., 79, 1086-1100.

Wakita, H. 1977. Normalization of vowels by vocal tract length and its application to vowel identification. IEEE Trans. Acoust. Speech Signal Process, ASSP-25, 183-192.