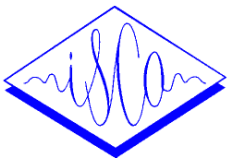


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

NORMALIZED TIME-FREQUENCY SPEECH REPRESENTATION IN ARTICULATION TRAINING SYSTEMS

Marcel Ogner and Zdravko Kacic

University of Maribor, Faculty of Electrical Engineering and Computer Science
WWW: http://www.dsplab.uni-mb.si

ABSTRACT

The aim of the work described in this paper is to develop and evaluate the speaker normalization technique based on the test to reference speaker mapping. The method is suitable for uniform time-frequency representation of speech used in speech corrector systems.

The normalized spectrum is generated after the analysis by synthesis for the given utterance using the MBE (multiband excitation) coding. The MBE speech production model decomposes the short time spectrum into the spectral envelope and excitation spectrum. The model offers the convenient way for joint vocal tract and excitation characteristics mapping to the reference speaker and at the same time preserving the phonetically relevant information in the test speaker utterance.

1. INTRODUCTION

In this paper we present the spectrum normalization technique we implemented in a teaching and training system for hearing handicapped children. The work is part of the INCO-COPERNICUS project SPECO (A Multilingual Teaching and Training System for Hearing Handicapped Children) with the goal of more uniform visual representation of speech characteristics to enable efficient training and accurate evaluation of correct pronunciation of vowels and fricatives.

The main source of inter-speaker variability in time-frequency representation of speech is due to the differences in vocal tract length and vocal excitation characteristics. Both components appear at somewhat different scale in its spectrum, the fine scale structure is mostly due to the excitation, while the intermediate scale structure is due to the vocal tract transfer function, characterized by its resonant frequencies. A common technique to suppress the influence of the excitation spectrum is to smooth the spectrum by linear convolution or cepstral smoothing and preserving only the spectral envelope. For higher pitched speech, such as children's speech, the accurate estimation of spectral envelope is a difficult task as the spectrum is less frequently sampled and vocal excitation often interferes with the spectral envelope estimation, especially at lower frequencies. Nevertheless, by suppressing the excitation component important information about voicing, necessary at the vocal tract normalization stage, is lost, since it does not make sense to calculate formant frequencies based on unvoiced speech data.

Preserving the excitation information in the spectrum requires additional excitation spectrum normalization, since variability in the fine scale structure in the spectrum, e.g. pitch, causes variability in energy distribution among frequency bands.

In order to perform the vocal tract length normalization and excitation spectrum normalization simultaneously, the new speech model is needed that would enable us the spectrum decomposition into spectral envelope and excitation model. The model has to be capable of spectrum resynthesis after some sort of manipulation on both components. The idea comes from speech coders that operate in frequency domain. The speech model that satisfies all our requirements is the MBE (Multiband Excitation) speech coder developed by Griffin and Lim [1].

The paper is organized as follows. Section 2 gives the briefly review of MBE speech model commonly used for speech coding. The attention is concentrated on details that improve the accuracy of spectrum decomposition into the source and spectral envelope estimation. Section 3 provides method for excitation spectrum normalization based on the MBE model. Section 4 proposes the vocal tract normalization technique developed from spectral envelope warping by applaying the frequency warping function extracted from the formant patterns for both speakers, reference and test speaker. Section 5 describes the objective tests employed for the evaluation of the technique and presents the results.

2. SPECTRUM DECOMPOSITION

The MBE model represents an input speech as the multi-plication of spectral envelope and an excitation spectrum. The model parameters are: pitch ω_o, amplitude envelope vector parameter $\vec{\mathbf{A}} = \{A_1, A_2, ...A_l, ...A_{L(\omega_o)}\}$ and V/UV ratio for particular frequency band. It is assumed that amplitude parameters are harmonic samples of an underlying vocal tract envelope and are allowed to be unconstrained free variables chosen to minimize the MSE criterion (1), where S_W is the measured short-time spectrum and $\hat{S}_W$ is the synthetic entirely voiced spectrum.

$$\varepsilon(\omega_o, \vec{\mathbf{A}}) = \sum_{n=0}^{N-1} \left| S_w(n) - \hat{S}_w\left(n, \omega_o, \vec{\mathbf{A}}\right) \right|^2 \qquad (1)$$

The minimization procedure is then repeated in a closed loop for each pitch value in the pitch region of interest. For the best parameter candidates that assure the optimum estimate and

when only the voiced component of the excitation is considered, the normalized mean squared-error for particular harmonic band can be written as

$$\varepsilon_l(\omega_o) = \frac{\sum_{\Omega_l} \left| S_W(n) - A_l E_W^V(n, \omega_o) \right|^2}{e^{-M\omega_o} \sum_{\Omega_l} \left| S_W(n) \right|^2} \qquad (2)$$

$$\Omega_l = \{n; l\omega_o - \omega_o/2 \le n \le l\omega_o + \omega_o/2\}.$$

In the original MBE model this error represents the V/UV degree for each harmonic band, binary declared either as voiced or unvoiced according to the predetermined threshold. The estimator in (2) shows the pitch biasing, e.g., for longer pitch periods the error tends to lower values. The unbiasing factor $1/e^{-B\omega_o}$ represents the inverse of the exponential fit to the error versus pitch function obtained by applying the white random noise to the input of the MBE model. The voiced component of the excitation E_W^V is generated as train of window frequency responses

$$E_w^v(n) = \sum_{l=1}^{L} W(n) * \delta(n - l\omega_o). \qquad (3)$$

In our experiments only the total error for entire frequency range $\varepsilon = \sum_{l=1}^{L} \varepsilon(\omega_o)$ is considered for the vocal tract normalization purpose described in section 4. Speech frames, for which total error exceeds certain experimentally determined threshold, have sufficiently weak periodicity so that they can be excluded from further vocal tract normalization processing.

Because the fidelity of the synthesized short time spectrum is for us more important than low bit coding, we use synthesis technique in which each harmonic band in the excitation spectrum can contain both voiced and unvoiced energies. In this method we used a V/UV mixture function for a short time speech spectrum to indicate the degree of V/UV mixture as a function of frequency. For the optimal MBE model parameters we assume that the mean of the original and the synthetic spectrum is equal for the frequency interval Ω_l [4].

$$\sum_{\Omega_l} S_w(n) = \sum_{\Omega_l} A_l \left[(1 - m(l)) E_W^V(n) + m(l) E_W^{UV}(n) \right] \qquad (4)$$

The unvoiced component of the excitation E_W^{UV} is generated as the spectrum of periodic random noise obtained by summation of sinusoidal signals with the same amplitude but with random phase (whispered speech). Normalizing the noisy spectrum with the norm $\left\| E_W^{UV} \right\|_2^2$ (as a result the mean is constant and equal one in the interval Ω_l) than the frequency dependent mixture function becomes

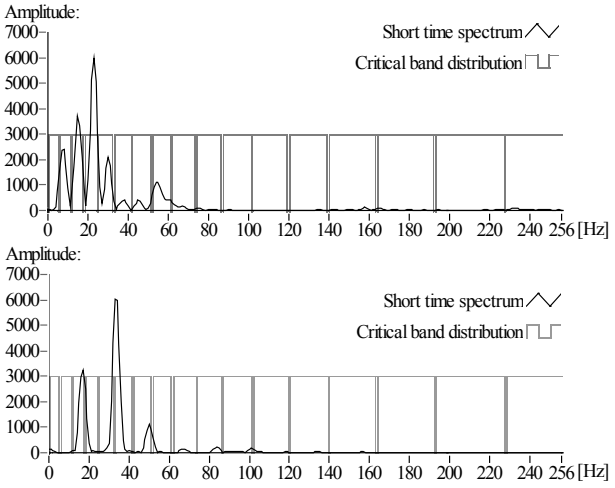
$$m(l) = \frac{\sum_{\Omega_l} \left[S_W(n) - A_l E_W^V(n) \right]}{\sum_{\Omega_l} A_l \left(1 - E_W^V(n) \right)}. \qquad (5)$$

3. EXCITATION SPECTRUM NORMALIZATION

In the joint time-frequency visual representation of speech the presentation of the phonetically relevant futures is of main importance, that is presenting the slow time-varying vocal tract transfer function estimated on a short-time basis. The common techniques such as cepstral smoothing or time-frequency filtering [3] obtain the smooth envelope of the power spectrum and suppress the excitation. These techniques suffer for accuracy when child speech is analyzed. On the other hand reducing the spectrum to the filter bank energy FBE, the excitation is lost, but the energy distribution among critical frequency bands is still excitation dependent (figure 1), especially at lower frequencies where the frequency bands are denser.

As the speech corrector system uses critical filter bank analysis as a preprocessor, we decided to preserve the spectrum with all its harmonic structure and rather normalize the excitation. The normalization of the excitation spectrum is carried out by first computing the MBE model parameters for both reference and test speaker:

- The voiced part of the excitation is synthesized by taking pitch from the reference speaker.
- The unvoiced (noisy) part of the excitation is synthesized with the same realization of the periodic random noise used for generating the excitation for reference speaker.
- Both components are mixed together according to the frequency dependent mixture function estimated for the test speaker.



Picture 1: Short time spectrums for the same sound /u/; a) male speaker, b) child speaker.

4. SPECTRUM ENVELOPE WARPING

In this section we describe a vocal tract normalization technique based on spectral shifts in the auditory filter domain. Because the final goal is visualization of the speech production rather then automatic speech recognition we permit to ourselves somewhat more freedom. Common way to extract acoustic features from speech is to obtain the smoothed estimate of the formant envelope. Further improvement can be obtained by mapping the real frequency scale (Hz) to perceived frequency scale (mel) or even more commonly, by computing equal-loudness weighted total energy only in critical bands around mel frequencies using critical band filters. Visualizing these features (spectrogram, cohleogram) will poses some variability, since different speakers have different formant frequencies for the same vowel, even if it is excellent pronounced. A main source for this variability among different speakers is due to the differences in vocal-tract lengths.

Conventional speaker normalization techniques use parametric approach and attempt to estimate constant scale factor between different speaker populations. In present study we use formant-based approach [2]. Let we say that reference speaker has formant pattern Fr={f_1, f_2, f_3,...,f_L} for a given utterance. Only the first three formants are considered for each frame. Frames with weak periodicity are discarded from further processing (section 2). For reference speaker we adopt usual mel critical band distribution, shown in picture 2, where each formant falls in certain weighted region of particular critical band filter. The patient using this articulation training system will produce somewhat different formant pattern for the same utterance. Moving his/her formants along mel-axis for each of N observations (it is assumed that frames are perfectly time aligned) gives us L points which coincide with formants of the reference speaker according to the filter index and its weighting on mel scale. But we still don't know how to shift critical band filters centered at mel frequencies to achieve formant matching. We solve this problem by polynomial fit. The model for polynomial fit is

$$y_i = \sum_{j=0}^{k-1} b_j x_i^j = b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_{k-1} x_i^{k-1} \quad . \qquad (6)$$
$$i = 0,1,2,\ldots..N-1 \;;\; k < N$$

The fitting problem is reduced the problem of finding coefficients B={b_1, b_2,..., b_{k-1}} that minimizes the difference between the observed data y_i and the predicted value. We use the least chi-square plane method to obtain coefficients in (6), that is, finding the solution B, which minimizes the quantity

$$\chi^2 = \sum_{i=0}^{N-1} \left(\frac{y_i - \sum_{j=0}^{k-1} b_j x_i^j}{\sigma_i} \right)^2 = \left| \mathbf{HB} - \mathbf{Y} \right|^2 , \qquad (7)$$

where **H** is observation matrix.

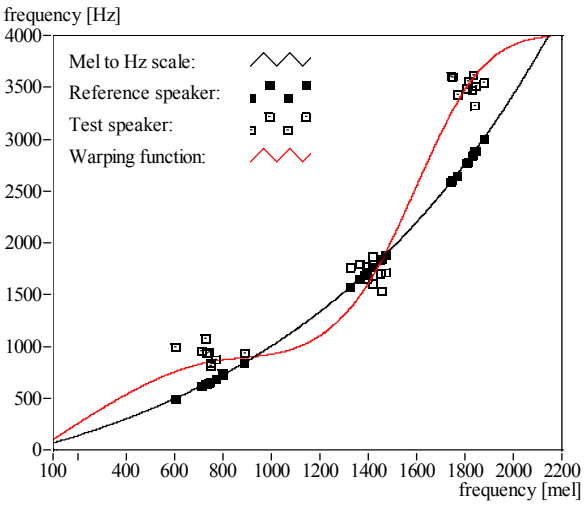
$$\mathbf{H} = \begin{vmatrix} 1 & f_0 & f_0^2 & . & . & f_0^{k-1} \\ 1 & f_1 & f_1^2 & . & . & f_1^{k-1} \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & f_{N-1} & f_{N-1}^2 & . & . & f_{N-1}^{k-1} \end{vmatrix} . \qquad (8)$$

If the formants are independent and normally distributed with constant variance, $\sigma_i = \sigma$ the preceding equation is also the least square estimation. One way to minimize χ is to set the partial derivatives of χ to zero with respect to b_1, b_2,..., b_{k-1}, which leads to matrix notation

$$\mathbf{H}^T \mathbf{HB} = \mathbf{H}^T \mathbf{Y} \; . \qquad (9)$$

Equation (9) can be solved using LU or Cholesky factorization algorithm. Subtracting modeled function from the reference gives us a warping function that determines necessary shift value at each frequency.

The drawback of this technique is that normalization works also when pronunciation is poor, so in that case algorithm also tries to reduce the distance between poor pronunciation of test speaker and correct articulation of reference speaker. From that reason certain local constraint has to be put on envelope warping.



Picture 1: Formant frequencies for 10 observations, presented in mel/Hz plane. Obtained error function determines necessary shift value at each frequency.

5. RESULTS

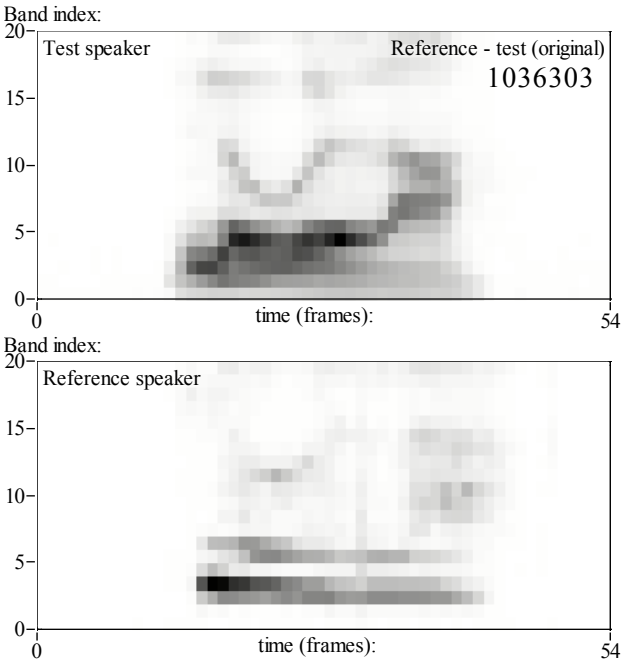
It's difficult to evaluate the efficiency of the method because of its specificity. We compared spectral distance between test and reference speaker before and after the normalization and for poor and for excellent pronunciation from the test speaker. In our experiments we expected that the distance for correct pronunciations after the normalization would be as small as possible and for poor articulation would stay practically unchanged.

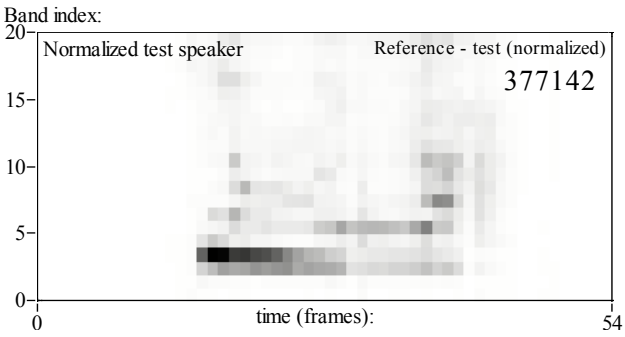
The database has been comprised of thirty speakers: male, female and children, ten speakers from each group. The words used were Hungarian isolated digits. The table 1 shows the averaged spectral distance decrease expressed in percents after the normalization takes place. As it can be seen the decrease is approximately 50 % in the case male to child and 40% in the female to child case.

The poor pronunciation has been simulated with incorrect pronunciation by inserting the randomly chosen wrong words instead of poor test speaker articulation, because we didn't have proper database of speech handicapped children. The reported values are absolute, because the distances were sometimes higher and sometimes lower after the normalization and are about 10%.

test/reference	word	correct [%]	incorrect [%]
male/child	/0/	55.9	9.1
	/1/	55.4	7.6
	/2/	38.8	11.5
	/3/	46.6	7.1
	/10/	52.1	6.1
female/child	/0/	49.2	7.1
	/1/	44.3	8.0
	/2/	34.4	9.8
	/3/	35.8	7.4
	/10/	40.0	4.9

Table 1: Weighted spectral decrease after the normalization for correct and incorrect pronunciation.





Picture 2: Spectrograms for the word /nula/; frequency resolution 20 bands. From top to bottom: male test speaker, reference child speaker, normalized test speaker.

6. SUMMARY

In this paper we presented the spectrum normalization technique suitable for uniform visual presentations of speech characteristics for audio-visual articulation training systems. The spectrum was decomposed using MBE model and resynthesized after excitation spectrum normalization and spectral envelope warping. The experiments were conducted by spectral distance comparison before and after the normalization. The results showed the significant distance decrease after the normalization while for the incorrect pronunciation the spectral distance remained in acceptable limits.

7. REFERENCES

1. D. Griffin and J.S. Lim, "*Multiband Excitation Vocoder*," in IEEE *Trans. Acoust., Speech and Signal Proc.*, ASSP-36, (8), 1988, pp. 1223-1235.

2. M. Ogner and Z. Kacic, "*Speaker Normalization for Audio-Video articulation Training*," Proc. ESCA Eurospeech99, pp. 579-582, 1999.

3. M. D. Riley, "*Speech Time-Frequency Repres-entations*," Kluwer Academic Publisher, Boston, 1989.

4. C. Laflamme, R Salami, R. Matmti, J-P. Adoul, "*Harmonic-Stochastic Excitation (HSX) Speech Coding Below 4 Kbit/s*," proc. of ICASSP, pp. 204-207, 1996.

5. R. J. McAluay and T. F. Quatieri, "*Speech Analysis/Synthesis Based on a Sinusoidal Representation*," IEEE *Trans. Acoust., Speech and Signal Proc.*, ASSP-34, (4), 1986, pp. 744-754