A NEURAL NETWORK FOR BLIND IDENTIFICATION OF SPEECH TRANSMISSION INDEX

Francis F. Li

Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, M1 5GD, UK
f.li@mmu.ac.uk

Trevor J. Cox

School of Acoustics and Electronic Engineering
Salford University,
Salford, M5 4WT, UK
t.j.cox@salford.ac.uk

ABSTRACT

A hybrid neural network model is proposed to determine the Speech Transmission Index of a transmission channel from transmitted speech signals without resort to prior knowledge of original speech. It comprises a Hilbert transform pre-processor, a PCA network for speech feature extraction and a multi-layer back-propagation network for non-linear mapping and case generalization. The developed method utilizes naturally occurring speech signals as probe stimuli, reduces measurement channels from two to one and hence facilitates speech transmission channel assessments under in-use conditions.

1. INTRODUCTION

Intelligibility is the most important concern of speech transmission channels, including acoustic ones such as rooms and electronic ones such as telephone lines, public address systems, codecs, etc.. The degradation of speech intelligibility is caused by the envelope shaping effect and additive noises of transmission channels. Such effects can be described by the Modulation Transfer Function (MTF) [1]. Derived from the MTF, Speech Transmission Index (STI), a single index defined to have good correlation with subjective perception of intelligibility, has been incorporated into international standards to quantify speech intelligibility of transmission channels [2,3]. In the standard STI method, the MTF is first identified using artificial test signals (sine-wave modulated white noises) and the STI is subsequently obtained by a series of linear and non-linear processing of the MTF data. Measurement of the STI is indeed a system identification and non-linear mapping problem.

It is well appreciated in acoustic research community that the use of artificial test signals in the STI method hinders in-use measurements. Endeavors have been made to accurately determine the STIs and other acoustic parameters from naturally occurring speech [4,5]. These methods rely on knowledge of speech signals at both input and output ends of a transmission channel and therefore are bi-channel measurements in nature. They cannot be used when the input signal is unknown. This paper presents a neural network model to extract the STI from transmitted speech signals without monitoring the original ones- a single channel blind system identification approach.

MTFs and STIs can be coarsely estimated by comparing envelopes of speech signals at transmitting and receiving ends [6], and better accuracy is achievable using a neural network compensation mechanism [7]. The challenge of the one channel approach is to estimate the speech envelopes at the input end of a channel from signals received at the output end. Principal components of received speech envelope signals are found robust to characteristics of transmission channels, giving useful profiles of envelopes of original speech signals. Unsupervised Principal Components Analysis (PCA) networks are adopted to identify envelope profiles of input speech from the outputs of transmission channels. Combining an Hilbert transform envelope detector, a spectrum estimator, a PCA network and a supervised back-propagation network, a hybrid model for blind STI identification is formed. The model is trained on and validated by a large number of examples of acoustic channels. Hypothetically, it should be applicable to various different channels.

2. NATURE OF STI IDENTIFICATION

STI is a single index derived from the MTF of a speech transmission channel [4]. A noise carrier *n(t)* is multiplied by a modulation function

$$m(t) = \sqrt{1 + m\cos(2\pi F t)} \qquad (1)$$

to generate an excitation signal

0-7803-7663-3/03/$17.00 ©2003 IEEE

$$i(t) = n(t) \cdot \sqrt{1 + m\cos(2\pi Ft)} \qquad (2)$$

where *F* is the modulation frequency and *m* is the modulation index. The intensity of excitation and response can thus be written as

$$I(t) = \mathbf{I}i[1 + m\cos(2\pi Ft)] \qquad (3)$$

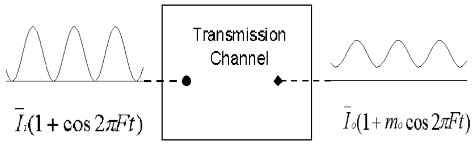
and

$$O(t) = \mathbf{I}o[1 + m_0 \cos 2\pi F(t - \varphi)] \qquad (4)$$

where *mo* is the modulation index of the output intensity function and φ is time delay due to transmission. ***I, Ii*** and ***Io*** are amplitudes of corresponding sinusoidal function (mean intensities). The MFT of a channel is defined as the ratio of *mo* to *m* as a function of modulation frequencies.

$$MTF(F) = \frac{m_o}{m} \qquad (5)$$

Figure 1. An illustration of MTF measurement



(for convenience, m=1)

For a good speech intelligibility, the envelope of speech signals should be well-preserved and noise interference minimized. The MTF describes the envelope shaping effect and takes noises into account, since input and output signal intensities are considered. Therefore it is closely correlated with intelligibility of transmitted speech. STI is a single index calculated from 98 MTF values at 14 1/3-octave band modulation frequencies (0.63, 0.80, …, 10.0, 12.5Hz) in seven octave bands (125, 250, …, 4k, 8kHz) in line with the following procedure [3]:

1. Converting *MTF(F)* into apparent *S/N* ratio

$$(S/N)_{app,\,F} = 10\log\left(\frac{MFT(F)}{1 - MFT(F)}\right) \qquad (6)$$

2. Limiting dynamic range to 30 dB
if (S/N)app > 15 dB >> (S/N)app = 15 dB
if (S/N)app < -15 dB >> (S/N)app= -15 d (7)
else (S/N)app=S/Napp

3.Calculation of mean apparent S/N ratio:

$$\overline{(S/N)}_{app} = \frac{1}{14}\sum_{F=0.63}^{12.5}(S/N)_{app},F \qquad (8)$$

4. Calculation of overall mean apparent *S/N* by weighting the *(S/N)app,F* of 7 octave bands

$$\overline{(S/N)}_{app} = \sum w_k \overline{(S/N)}_{app,\,F} \qquad (9)$$

where *wk*=0.13, 0.14, 0.11, 0.12, 0.19, 0.17 and 0.14 respectively for the 7 octave bands .

5. Converting to an index ranging from 0 to 1

$$STI = \frac{\overline{(S/N)}_{app} + 15}{30} \qquad (10)$$

It is apparent that STI is a purpose defined parameter from the MTF and STI estimation is in fact a system identification problem. When extracting STIs from the transmitted signals, the problem becomes blind system identification.

3. THE HYBRID NEURAL NETWORK

3.1. Rationale

The MTF can be estimated from the envelope spectra of original and transmitted speech by [6]

$$MTF(F) \approx \frac{Ey(F)}{Ex(F)} \qquad (11)$$

where E_X *(F)* and E_Y*(F)* are the envelope spectra of input and output speech signals of a channel. When this is compared with the definition of the MTF, it becomes apparent the estimation errors stem from the difference between the actual speech signal and the sinusoidal modulated white noises. Estimation accuracy can be improved by introducing certain compensation mechanisms; in particular a neural network approach can be adopted [7]. To blind-identify MTFs from output signals solely, information about *Ex* becomes an additional challenge.

3.2. The hybrid model

The design consideration of the hybrid neural network is to use a good spectrum estimator to obtain speech envelope spectra, an unsupervised network to acquire input envelopes profiles from output signals and finally a supervised network to map the MTF onto STI, compensate spectral difference and perform case generalization. Figure 2 illustrates the framework of the proposed hybrid neural network system.

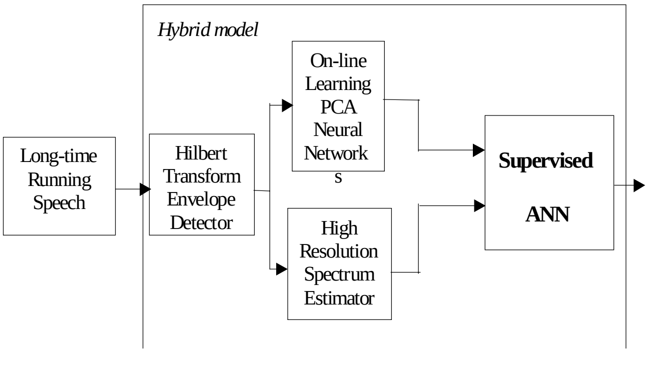

Hybrid model
Long-time Running Speech
Hilbert Transform Envelope Detector
On-line Learning PCA Neural Networks
High Resolution Spectrum Estimator
Supervised ANN

Figure 2. Hybrid neural network model

3.2.1. Hilbert transform envelope detector

The envelope *ev(t)* of a 60 second running speech excerpt *s(t)* is first determined by

$$ev(t) = \sqrt{s^2(t) + s_h{}^2(t)} \tag{12}$$

where *Sh(t)* is the Hilbert transform of *s(t)*

$$s_h(t) = H[s(t)] \equiv \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{s(t-t')}{t'}dt' \tag{13}$$

The envelope spectra are further estimated by Welch's average periodogram with a Hanning window. Empirical comparisons revealed a higher resolution and more data points than defined in the standard STI method benefit the accuracy of the blind identification. Envelope spectra from immediately above DC to 25Hz at 0.5 Hz intervals are used.

3.2.2. Input speech profiling

Profiles of input signal envelopes is obtained using a PCA neural network. An m-tap delay-line and a rectangular observation window is first applied to accomplish the necessary conversion from received speech envelope signals to a multi-dimensional data space for PCA as depicted in Figure 3. The speech envelope, low-pass filtered and decimated to 40 samples/s, is passed through the delay-line and then windowed to obtain m-dimensional observations. A 125-500ms window is empirically found to be appropriate. Each column in the data space (reconstruction space) forms one observation of the envelope signal. The reconstruction space is used to train a PCA neural network shown in Figure 4.

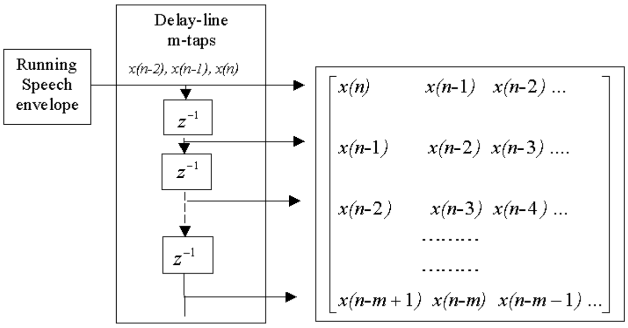

Running Speech envelope
Delay-line m-taps
x(n-2), x(n-1), x(n)
z^{-1}
z^{-1}
z^{-1}
$x(n)$ $x(n-1)$ $x(n-2)$ ...
$x(n-1)$ $x(n-2)$ $x(n-3)$ ....
$x(n-2)$ $x(n-3)$ $x(n-4)$ ...
....
....
$x(n-m+1)$ $x(n-m)$ $x(n-m-1)$ ...

Figure 3. Reconstruction space envelop signal

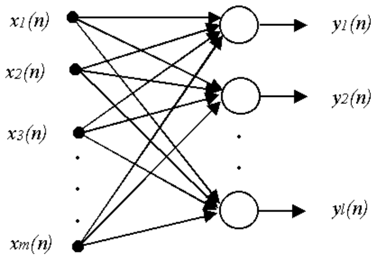


Figure 4. PCA network

The network has *m* inputs and *l* outputs (*l*<*m*). The trainable parameters are synaptic weights w_{ij} which connects the i^{th} input to j^{th} neuron, where *i=1,2,..m and j=1,2,..l.* The dynamic equation of such a network is

$$y_j(n) = \sum_{i=1}^{m} w_{ji}(n)x_i(n) \tag{14}$$

and the generalized Hebbian algorithm [9]

$$\Delta w_{ij}(n) = \eta\left[y_j(n)x_i(n) - y_j(n)\sum_{k=1}^{j} w_{ki}(n)y_k(n)\right] \tag{15}$$

is applied to perform unsupervised learning. The weight w_{ij} of neuron *j* converges to i^{th} component of the eigenvector related to the j^{th} eigenvalue of the correlation matrix. The outputs are the first l^{th} maximum eigenvalues. It is informative to observe the first principal component of speech envelopes through different acoustic transmission channels-see Figure 5.

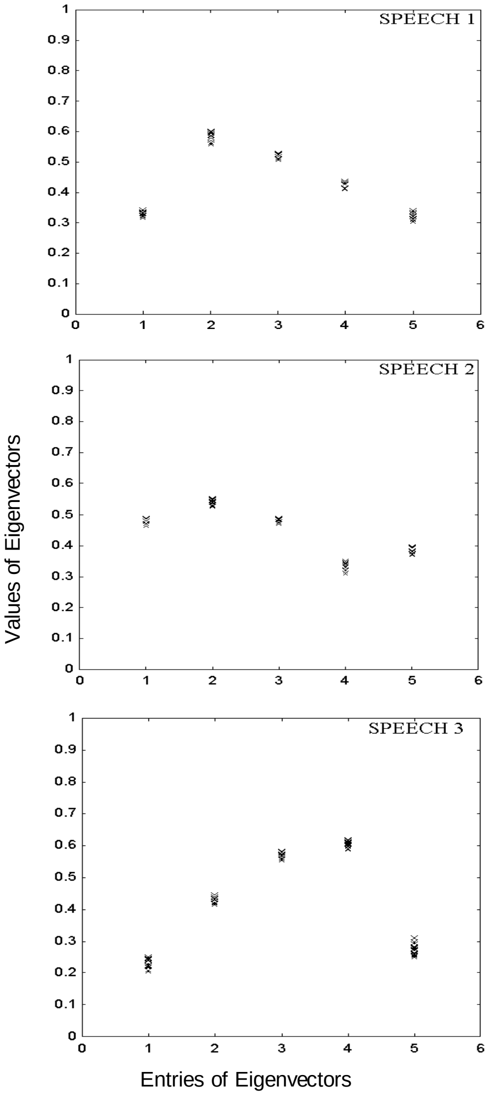


Figure 5. Over plot of eigenvectors

In this experiment, three different speech signals transmitted through 50 different acoustic channels with different reverberation times from 0.1 to 5s are observed. Three graphs show the over-plots of eigenvectors of the first principal components for 50 different reverberation times, where the y-axis is the value of entries of eigenvectors and x-axis indicates entry number of the eigenvectors. The results show that different input speech signals have distinctive eigenvectors, while characteristics of transmission channels only affect them slightly. Eigenvectors of principal components give robust profiles of speech envelopes!

Eigenvectors of the first two principal components are fed into the supervised final stage to give information about the envelopes of input speech signals, while the eigenvalues are discarded. The final stage is a typical back-propagation network, with two non-linear hidden layers and only one linear neuron on the output layer. Linear basis and sigmoid activation functions are used for hidden layers, and training follows standard back-propagation algorithm [9]. Thus a hybrid neural network model is formed.

4. TRAINING AND VALIDATION

The neural network model has a supervised final stage, therefore training examples are needed. A large number of impulse responses of acoustic transmission channels are used and convolved with 18 different anechoic running speech excerpts to generate training examples. Since STI method takes background noises into account, white noises are assumed in training and validation phases. Teacher values are obtained from impulse responses of transmission channels and noise levels via the standard routine as outlined in Section 2. Eigenvectors of first and second principal components obtained by the PCA sub-network and the envelope spectra described before are fed into supervised neural network. So in both training and retrieve phases of the hybrid model, the unsupervised learning to obtain the PCA values is performed. The trained hybrid network is tested with acoustic and speech cases not seen in the training phase. The maximum prediction errors found in the full range ($0 \leq \text{STI} \leq 1$) test is 0.087. The results indicate that the proposed method can usefully blind-identify STIs from transmitted speech signals, but not as accurately as when the standard method is used.

5. CONCLUSION AND DISCUSSIONS

A hybrid neural network model to perform blind identification of STI for speech transmission channels is developed and validated via simulations. Acoustic transmission channels are considered in the presented paradigm, but the method should also be applicable to electronic transmission channels showing envelope shaping effect on speech signals. In electronic transmission channels, however, system non-linearities are common. Nevertheless, this hybrid model should inherently be able to deal with this effect, provided there are suitable examples for training. Moreover, It is worth noting that the prediction accuracy may be further improved by optimized design of each of the building blocks in the hybrid model.

From an application standing point, the proposed method should facilitate in-use measurements of STI and resolve many dilemmas in site measurements. Moreover, identifying characteristics of speech transmission channels is often the first step of implementing inverse filters for channel equalization. Blind identification with naturally occurring signals are particularly useful when time variance is non-trivial, since it enables monitoring the channel when in-use and so adapting the inverse filter in real-time.

5. REFERENCES

[1] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica, Vol. 28, pp. 66-73, 1973.

[2] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech transmission quality,” J. Acoust. Soc. Am. Vol. 67, No. 1, pp. 318-326, 1980.

[3] IEC Standards, 60268-16:1998, (also BS EN 60268-16 and BS 60268-16), Sound system equipment, Part 16: Objective rating of speech intelligibility by speech transmission index, 1998.

[4] F. Li and T. J. Cox, “Predicting speech transmission index from speech signals using artificial neural networks,” Proc. World Multi-conference on Systemics, Cybernetics and Informatics, SCI 2000’, Florida, Vol. 6-II, pp. 43-47, 2000.

[5] T. J. Cox, F. Li, and P. Darlington, “Extraction of room reverberation time from speech using artificial neural networks”, Journal of AES, Vol. 49, No. 4, pp. 219-230, 2001.

[6] H. J. M. Steeneken and T. Houtgast, "The temporal envelope spectrum of speech and its significance in room acoustics," 11th ICA conference publication, Paris, 1983.

[7] T. J. Cox and F. F. Li, "Using Artificial Intelligence to Enable Occupied Measurements of Concert Hall Parameters," Proc. 17th ICA, Italy, 3A.08.05., 2001.

[8] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," Neural Networks, Vol. 12, pp. 459-473, 1989.

[9] D. E. Rumelhart, G. E. Hinton, G. E. and R. J. Williams, “Learning internal representations by error propagation,” in Parallel distributed processing: Exploration in the Microstructure of Cognition, MIT Press, Cambridge, MA., vol.1, pp. 318-362, 1986.