BLIND CHANNEL IDENTIFICATION IN SPEECH USING THE LONG-TERM AVERAGE SPEECH SPECTRUM

Nikolay D. Gaubitch, Mike Brookes and Patrick A. Naylor

Imperial College London, Exhibition Road, SW7 2AZ, UK
{ndg,mike.brookes,p.naylor}@imperial.ac.uk

ABSTRACT

Estimation of the magnitude response of an unknown channel in single-microphone speech signals is considered. It is shown how the Long-Term Average Speech Spectrum (LTASS) can be used to identify the unknown channel and a blind channel identification algorithm is developed based on that. Furthermore, an established approximate formula for LTASS is demonstrated to be a useful tool in the context. The algorithm is evaluated using a weighted spectral distortion measure using simulated, measured and real channels with various distinct spectral characteristics. It is demonstrated that the algorithm can identify accurately the magnitude spectrum of an unknown channel in noise-free conditions. We also show results for three different additive noises where estimation accuracy is reduced but the degradation varies largely, depending on the long-term spectral characteristics of the noise.

1. INTRODUCTION

When a speech signal is captured with a microphone and stored or transmitted prior to being observed by a listener, inevitably it is altered by the acoustic medium, the microphones and the storage/transmission medium. These alterations often have detrimental effects on the quality and/or intelligibility of the observed speech signal.

Consider a speech signal, $x(n)$, observed by a listener at a different location to that of the talker. The observed signal consists of the desired speech signal, $s(n)$, with its spectral characteristics modified, for example, by an acoustic channel or a microphone whose effects are characterized by an impulse response $h(n)$. There will also be some additive measurement noise, $v(n)$. The relationship between speech, channel and noise at the point of observation can be expressed as

$$x(n) = s(n) * h(n) + v(n), \tag{1}$$

where $*$ denotes linear convolution. The work presented in the following sections will be developed in the frequency domain. It is customary to process speech in the frequency domain using short overlapping frames and, accordingly, the expression in (2) can be written

$$X_l(k) = S_l(k)H_l(k) + V_l(k), \tag{2}$$

where $X_l(k)$, $S_l(k)$, $H_l(k)$ and $V_l(k)$ are the Short-Time Fourier Transforms (STFT) of the lth frame of $x(n)$, $s(n)$, $h(n)$ and $v(n)$ respectively and k is the frequency bin index.

Different channels and different noise types have different spectral and statistical properties. Depending on these properties, they can reduce the perceived quality and, sometimes, the intelligibility of the observed speech signal. Estimation and reduction of additive noise is a widely researched topic and there are numerous algorithms available in the literature [1, 2]. In this paper, we consider estimation of the magnitude response of the unknown channel, $H_l(k)$. The method which we present uses the Long-Term Average Speech Spectrum (LTASS). A related idea was presented in [3] where the authors studied channel equalization in the context of speaker recognition. The material presented here differs in a number of ways: (i) we use an iterative approach for the estimation of the average spectrum of the observed signal rather than the average of the total observation; (ii) we investigate the use of two alternatives for LTASS; (iii) we provide experiments with three different additive noises; (iv) we evaluate the algorithm using an objective measure of channel similarity.

The remainder of the paper is organized as follows: In Section 2, we discuss the principle and the implementation of the channel identification algorithm. In Section 3, we introduce a weighted spectral distortion measure for evaluation of the estimated channels and use it to evaluate the algorithm's performance for a variety of channels. Finally, we draw conclusions from this work in Section 4.

2. BLIND CHANNEL IDENTIFICATION

In this Section, we derive the blind channel identification method. First the principle behind the method is described and the effects of noise are discussed. Then, details of the implementation are described.

2.1 Principle

The objective is to estimate the magnitude response of the unknown channel. Therefore, we consider the power spectrum of the observed speech signal in (2)

$$\begin{aligned} |X_l(k)|^2 &= |S_l(k)|^2|H_l(k)|^2 + |V_l(k)|^2 \\ &\quad +2|S_l(k)||H_l(k)||V_l(k)|\cos(\angle\Delta), \end{aligned} \tag{3}$$

where the angle $\angle\Delta = \angle S_l(k) + \angle H_l(k) - \angle V_l(k)$.

Next, it is assumed that the signal and the noise magnitudes and phases are independent, the channel is stationary or varies much slower than the speech and $E\{\cos(\angle\Delta)\} = 0$, where $E\{\cdot\}$ denotes expectation. Then, taking the expecta-

tion on both sides of (3), we get

$$\begin{aligned} E\{|X_l(k)|^2\} &= E\{|S_l(k)|^2\}|H(k)|^2 + E\{|V_l(k)|^2\} \\ &= E\{|S_l(k)|^2\}\left(|H(k)|^2 + \frac{E\{|V_l(k)|^2\}}{E\{|S_l(k)|^2\}}\right) \\ &= \bar{P}_S(k)\left(P_H(k) + \frac{\bar{P}_V(k)}{\bar{P}_S(k)}\right) \\ &= \bar{P}_X(k), \end{aligned} \quad (4)$$

where $P_X(k)$ denotes the power spectrum of a signal $x(n)$ and $\bar{P}_X(k)$ the expected power spectrum of a signal $x(n)$. $\bar{P}_S(k)$ in (4) is the mean value of the short-term speech spectrum, which can be approximated by the LTASS. Therefore, we can estimate the log spectrum of the channel using

$$\log\left(P_H(k) + \frac{\bar{P}_V(k)}{\bar{P}_S(k)}\right) = \log(\bar{P}_X(k)) - \log(P_{\text{LTASS}}(k)), \quad (5)$$

where $P_{\text{LTASS}}(k)$ is some predefined model of the LTASS which, in practice, can be found either by measurement [4] of many talkers and many utterances or by using an approximate formula [5]. A comprehensive study in [4] shows that LTASS is relatively invariant with language but there are differences between male and female talkers, particularly at lower frequencies. Nevertheless, it is possible to find a reasonable average representation for both sexes.

In the noise-free case, $v(n) = 0$, the component on the left hand side of (5) $\bar{P}_V(k)/\bar{P}_S(k)$ equals zero. The channel can then be identified to an accuracy depending on the accuracy of the assumed LTASS model

$$\log(\hat{P}_H(k)) = \log(P_H(k)) + \varepsilon(k), \quad (6)$$

where $\varepsilon(k) = \log(\bar{P}_S(k)/P_{\text{LTASS}}(k))$ is an error due to the discrepancy between the assumed and the actual LTASS. A second factor in the accuracy of this estimate is the length of the available speech signal such that $\bar{P}_X(k)$ can be estimated accurately. The channel properties will not have as significant an effect as the speech signal characteristics.

When noise is present, there is an additional error in the estimate. The actual impact of this error will depend on the long-term average spectrum of the noise, the channel characteristics and on the noise power ($\bar{P}_V(k)/\bar{P}_S(k)$ in (5) is inversely proportional to the SNR). Both the noise-free and the noisy cases will be discussed further with the simulation examples in Section 3.

2.2 Implementation

In order to implement an algorithm based on the ideas described in Section 2.1, we need two components: (i) a model for the LTASS and (ii) a procedure for calculating the average spectrum of the observed signal. These two components are discussed in the following.

There are two possible approaches to obtaining an LTASS model and we consider both. In the first approach, the LTASS can be found by measurement from many talkers and many utterances [4]. We used the complete training set of the TIMIT database to extract the LTASS as an average from the short-term spectra of all utterances and all talkers. The training set of TIMIT contains anechoic noise-free recordings from 422 male and 184 female talkers, with ten

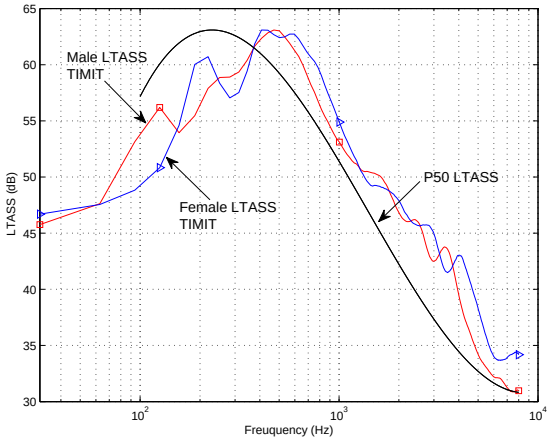

Figure 1: Male and female LTASS measured from TIMIT and generic LTASS from P50 calculated with (7).

sentences from each talker; the duration of each sentence is approximately three seconds. Frames of 16 ms and with 50% overlap were used for the STFT and male and female talkers were processed separately.

Alternatively, an approximate formula can be used to approximate the LTASS. One such formula is defined in the ITU-T recommendation P.50 [5] to

$$\begin{aligned} P_{LTASS}(f) = &-376.44 + 465.439\log_{10} f \\ &- 157.745(\log_{10} f)^2 + 16.7124(\log_{10} f)^3 \text{ dB}, \end{aligned} \quad (7)$$

where f is frequency in Hz and the output is a log intensity spectrum in dB relative to $1W/m^2$.

The LTASS from (7) and the LTASS calculated from the TIMIT database for male and female talkers are shown in Figure 1. It can be seen that the measured male and female LTASS mostly differ in the low frequency region. The overall shape of the measured LTASS is very similar to the results shown in [4]. We use an average of the male and female LTASS in the blind channel identification algorithm. We also see that, the LTASS calculated from (7) differs from the LTASS measured from TIMIT in the peak frequencies, which can reduce the estimation accuracy. Nevertheless, the advantage of having a formula to approximate the LTASS is that of straightforward reproducibility of the algorithm.

Next, we consider the estimation of the average spectrum of the observed signal. In order to account for channel variability and to accommodate frame-by-frame processing, we implement the averaging using exponential smoothing according to

$$\log(\hat{P}_H(k,l)) = \alpha\log(\hat{P}_H(k,l-1)) + (1-\alpha)\hat{H}(k,l) \quad (8)$$

where

$$\hat{H}(k,l) = \log(P_X(k,l)) - \log(P_{\text{LTASS}}(k)) \quad (9)$$

is the instantaneous estimate of the channel magnitude response in frame l and $0 \leq \alpha \leq 1$ is the smoothing factor. Both $P_X(k,l)$ and $P_{\text{LTASS}}(k)$ are normalized using log-spectral

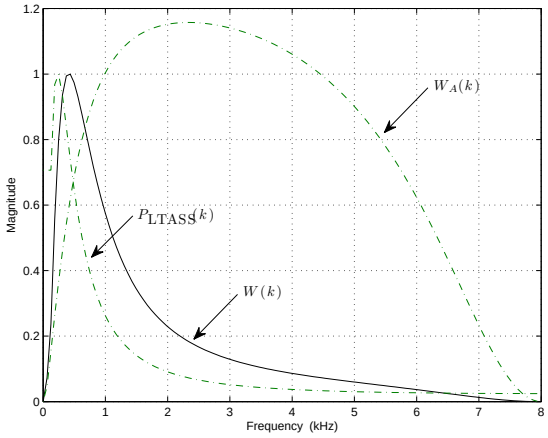

Figure 2: Composite weighting function (solid) used in the weighted spectral distance defined in (11) and its two components (dashed).

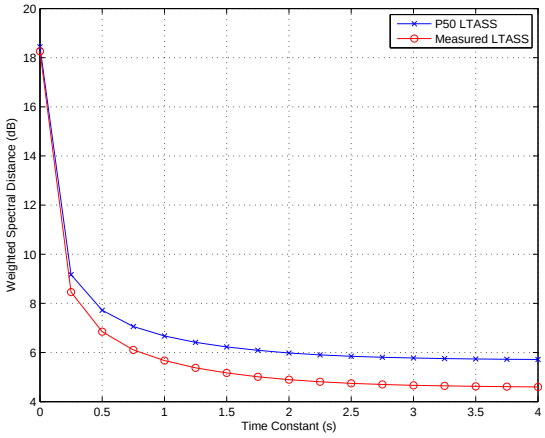

Figure 3: Time-constant versus channel identification performance in terms of weighted spectral distance.

mean subtraction prior to their application in (9), in order to avoid any arbitrary scale factor issues. The choice of α is studied in Section 3 and it is set using the relation to the time constant given by

$$\alpha = \frac{\tau}{(\tau + T_S)}, \tag{10}$$

where T_S is the sampling period and τ is the time constant which, in this case, defines the time for the channel estimation in (8) to reach a steady-state value.

3. EXPERIMENTS AND RESULTS

We now present a performance evaluation of the algorithm described in Section 2. We will introduce a metric used for the evaluation of an estimated channel spectrum and we will use this in a series of experiments to demonstrate various aspects of the LTASS-based blind channel identification algorithm.

3.1 Evaluation

We consider a weighted error measure in order to compare two power spectra $P_1(k)$ and $P_2(k)$. A weighted root mean squared log-spectral distance can be defined as [6]

$$d(P_1,P_2,k) = \left[\frac{\sum_{k=0}^{N-1} W(k)|e(k)|^2}{\sum_{k=0}^{N-1} W(k)}\right]^{\frac{1}{2}} \text{dB}, \tag{11}$$

where

$$e(k) = 10\log_{10}\left(\frac{P_1(k)}{P_2(k)}\right), \tag{12}$$

and $W(k)$ is a frequency dependent weight function.

One of the key purposes of channel identification is to use the estimated channel response to neutralize its effect on the speech signal. Since we assume explicitly that the desired signal is speech and the destination is a human listener, a weighting function should feature some human speech production and hearing properties. A good weighting function which we propose here includes A-weighting, $W_A(k)$, and LTASS weighting and is defined as

$$W(k) = W_A(k)P_{\mathrm{LTASS}}(k). \tag{13}$$

The composite weight function, $W(k)$, and its two components $W_A(k)$ and $P_{\mathrm{LTASS}}(k)$ are shown in Fig. 2. Note that $W(k)$ has been normalized to unity.

3.2 Simulation Results

For the following illustrative experiments, data is drawn from the core test sets of the TIMIT database. The core test set (including the dialect sentences) consists of 240 sentences, ten sentences from each of the 16 male and 8 female talkers. In this way, we use different data from that used to estimate the LTASS. Simulated channel responses are generated by placing poles or zeros in conjugate pairs inside the unit circle; the magnitude of each pole or zero is restricted to the range $0.8-0.99$. The positions of the poles and the zeros are chosen randomly from a uniform distribution. All experiments are performed using both the measured LTASS from TIMIT and the approximate formula from (7). The weighted spectral distance is calculated for one utterance as the average across all frames after τ seconds.

3.2.1 Experiment 1: Time-Constant Selection

For the first experiment, we used a two-pole channel and no additive noise. The smoothing factor was set as in (10) by varying the time-constant, τ, in the range $0-4$ s in 0.25 s increments. The sentences from the test set of TIMIT were used by concatenating all sentences from each talker to form one longer sentence per talker of approximately 30 s, resulting in a total of 24 different sentences and 24 talkers. The channel is estimated for each utterance and for each time-constant. The results, averaged over all utterances, are shown in Fig. 3. Based on this result, the performance floor is reached at around $\tau = 2.5$ s; this is thus the chosen time-constant value we use for the remaining experiments. Choosing a lower time constant, would allow faster tracking of a

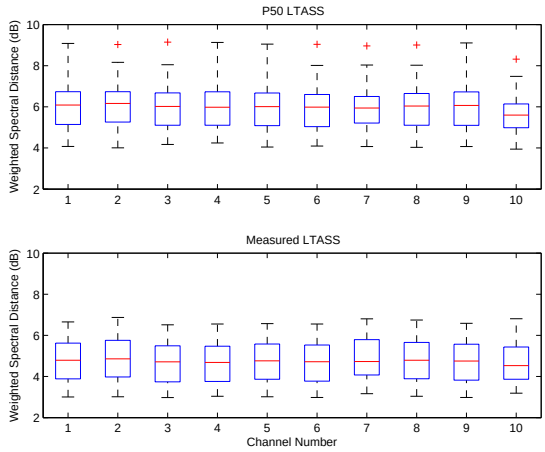

Figure 4: Ten different channels comprising one conjugate pair of poles and one conjugate pair of zeros and 48 utterances, two by each of 16 male and eight female talkers.

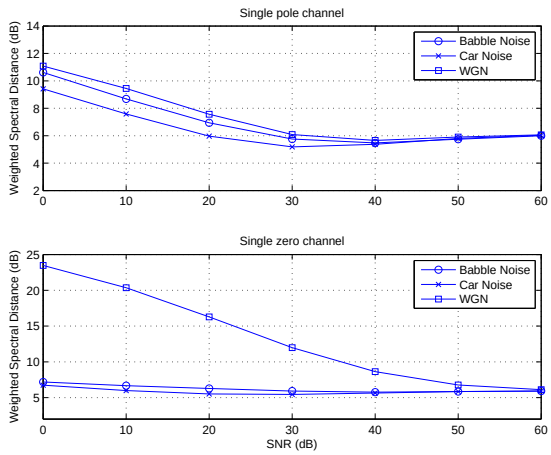

Figure 5: Channel estimation in noise with P50 LTASS.

time-varying channel at the cost of some estimation accuracy.

3.2.2 *Experiment 2: Channels vs. Speech*

In the second experiment we used ten randomly generated channels comprising one conjugate pole pair and one conjugate zero pair and there is no additive noise. The TIMT test sentences were concatenated using five sentences to create one utterance. This results in two utterances per talker and a total of 48 utterances. The results are shown in the box plot in Fig. 4. This plot shows that there can be a large variation in performance with different utterances while there is much less dependence on the channel. This indicates that longer-term averaging over utterances may improve performance, provided that the channel is stationary.

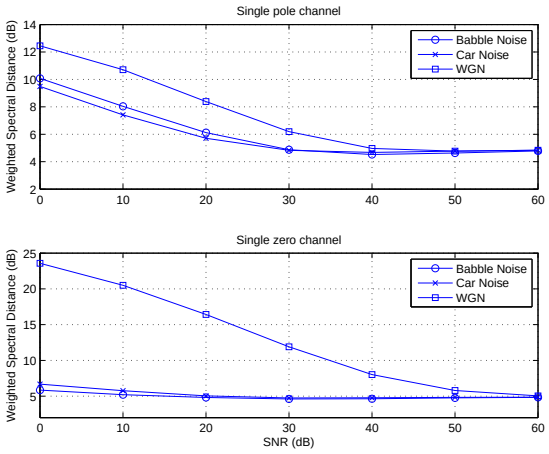

Figure 6: Channel estimation in noise with measured LTASS.

3.2.3 *Experiment 3: Identification in Noise*

Next, we used two fixed channels, one with a conjugate pair of poles and one with a conjugate pair of zeros. We then added noise to the filtered signals, varying the SNR between 0 dB and 60 dB. Three different types of noise were considered: babble noise, car noise and White Gaussian Noise (WGN). Figure 5 shows the channel identification results in terms of weighted spectral distance when using P50 LTASS and Fig. 6 shows the results obtained with measured LTASS.

One observation that is made is that the results have the same rank order in all cases, with the WGN resulting in the worst identification performance and car noise with the best. A possible explanation to this can be given by considering the term $P_V(k)/P_S(k)$ in (5). It is seen that if the long-term average of the noise equals that of LTASS then this term equals one at all frequency bins and the channel estimate will be as accurate as in the noise-free case. The long-term average of the WGN is flat, while babble noise and car noise have similar spectral trend as the LTASS with stronger magnitudes in the low frequency regions. Consequently, the impact of the car noise is much lower than the impact of the WGN. In practice, the relationship between the channel response and this inverse SNR term is more complex due to the averaging being performed in the log-spectral domain.

3.2.4 *Experiment 4: Real Measured Channels*

Finally, we show two illustrative examples with two real channels; the objective is to demonstrate the use performance of the algorithm with realistic data and to relate the numbers of the weighted spectral distance to quintessential estimation examples. First, a measured microphone response was convolved with clean speech. The true and the estimated channels are shown in Fig. 7. The weighted spectral distance in this example is 4.64 dB for the estimation using P50 LTASS and 3.87 dB for the estimation using measured LTASS. We see that the important large scale components (the position of the three poles in this case) have been identified correctly.

Secondly, we used a sample from NTIMIT, which is the TIMIT database recorded over real noisy telephone network channels. The database also provides measurements of the

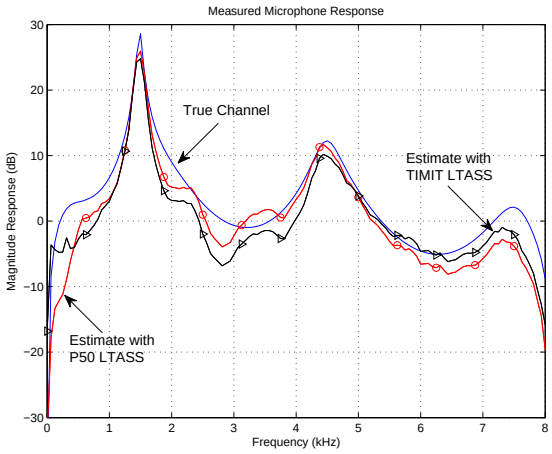

Figure 7: Measured microphone response and estimates of it in a noise-free case. The weighted spectral distance is 4.64 dB for the estimation using P50 LTASS and 3.87 dB for the estimation using measured LTASS.

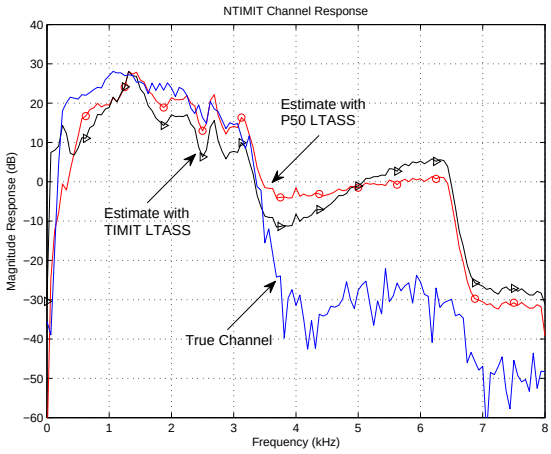

Figure 8: Measured channel of a noisy telephone network from NTIMIT. The weighted spectral distance for this is 33.34 dB for the estimation using P50 LTASS and 36.13 dB for the estimation using measured LTASS.

channel responses. The measured and the estimated channels are shown in Fig. 8. The weighted spectral distance for this is 33.34 dB for the estimation using P50 LTASS and 36.13 dB for the estimation using measured LTASS. We see from the graphs that the large estimation error can be attributed to the high frequency portion of the channel ($\geq$ 4 kHz). This is a combination of the speech signal in this portion of the spectrum being buried in noise and the relatively low energy in the speech signal in frequencies above 4 kHz as is evident from the LTASS in Fig. 1.

From this set of experiments, it can be seen that using the measured LTASS can result in a small improvement over the approximate LTASS equation from (7) in terms of weighted spectral distance. Therefore, it provides a useful tool for blind channel identification, particularly in terms of the dominant characteristics of the channel.

4. CONCLUSIONS

We have developed a method that uses a pre-defined model of the long-term average speech spectrum to blindly identify a channel that is either stationary or slowly varying. A weighted spectral distortion measure, suitable for speech signals, was proposed and employed to evaluate the algorithm using simulated channels, measured microphone responses and telephone transmission channels from the NTIMIT database. It was shown, for the noise-free case, that this method can identify the large scale features of a channel in a manner that works well for a wide variety of channels but its performance can vary with different speech utterances. In the presence of noise, it was found that the performance degrades differently depending on the spectral characteristics of the noise, on the noise power and on the channel. Finally, the potential of the method was demonstrated with real measured channel responses from a microphone and from a noisy telephone network showing that this is a promising method for identification of the larger scale features of a channel.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement Theory and Practice*. Taylor & Francis, 2007.

[2] J. Benesty, J. Chen, Y. Huang, and I. Cohen, Eds., *Noise Reduction in Speech Processing*. pub-SV:Berlin, Germany, 2009.

[3] S. J. Wenndt and A. J. Noga, "Blind channel estimation for audio signals," in *Proc. IEEE Aerospace Conf.*, vol. 5, 2004, pp. 3144–3150.

[4] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. E. Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, , T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," *Journal Acoust. Soc. of America*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[5] *Telephone Transmission Quality Measuring Apparatus: Artificial Voices*, International Telecommunications Union (ITU-T) Recommendation P.50, Mar. 1993.

[6] R. Viswanathan, J. Makhoul, and W. Russell, "Towards perceptually consistent measures of spectral distance," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1976, pp. 485–488.