Characterisation and simulation of telephone channels using the TIMIT and NTIMIT databases

Herman Kamper and Thomas Niesler

Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
{14819821,trn}@sun.ac.za

Abstract

The paper presents techniques that allow the effects of a variety of telephone channels to be simulated, given wideband speech recordings. By comparing corresponding utterances from the TIMIT and NTIMIT corpora, both a channel and a noise model were derived. These models were shown to closely mimic the spectral effects of the NTIMIT telephone network. The application of the techniques to the development of ASR systems indicates that the noise model is the major factor leading to an increased accuracy from a basic bandpass channel approximation.

1. Introduction

The development of telephone-based speech recognitions systems, such as automatic call routing and spoken dialogue systems, requires speech corpora that have been recorded over a variety of telephone channels. Other applications, such as dictation systems, are developed using speech recorded through high quality microphones. There is therefore a tendency to compile both narrowband telephone-quality as well as studio-quality wideband speech corpora for the development of speech applications. This doubles the effort involved in corpus development, which is particularly disadvantageous when dealing with under-resourced languages.

This paper describes the development of techniques that allow the effects of a variety of telephone channels to be simulated, in order to allow realistic telephone-quality speech to be obtained from wideband recordings. The telephone channel characteristics were estimated by comparing corresponding utterances from the well-known TIMIT and NTIMIT corpora.

The paper is structured as follows. A short description of the TIMIT and NTIMIT databases is given first. This is followed by a description of the chosen model of the telephone channel. Next, the techniques used to determine filter responses from the TIMIT/NTIMIT utterances are presented, followed by an analysis of the noise introduced by the telephone channel. Finally, the application of the techniques to the development of ASR systems is evaluated.

2. The TIMIT and NTIMIT databases

A thorough description of both the TIMIT and NTIMIT databases is given in [1]. The TIMIT database was compiled using 630 speakers of whom 438 were male and 138 female. Each speaker spoke ten utterances giving a total of 6 300. Utterances were transcribed both phonetically and orthographically.

The NTIMIT database was compiled by transmitting the TIMIT utterances over a physical telephone network. Great care was taken to ensure that the NTIMIT utterances accurately reflect the general nature of telephone-based speech. An acoustically isolated room, an artificial mouth and a telephone test frame (a device controlling the positioning of the artificial mouth relative to a telephone handset) was used. Speech was processed at a sampling frequency of 16 kHz and encoded as 16 bit PCM. A 9th order elliptical anti-aliasing filter with a cut-off frequency of 6.4 kHz was used in the sampling process.

The USA is divided into Local Access and Transport Areas (LATAs) which are geographical regions corresponding to the subdivision of the telephone network. Within each LATA different central offices, which handle calls, are found. A variety of telephone channels were used to collect the NTIMIT database by varying the central office (and thus the geographical location) to which each TIMIT utterance was transmitted. In total 253 central offices, and thus 253 different telephone channels, were used in the compilation of the NTIMIT corpus.

3. Model of the telephone channel

The model used to simulate the telephone channel is shown in Figure 1. The wideband input sequence $x[n]$, which in our case would correspond to TIMIT speech, is bandlimited by $\hat{H}(z)$, which simulates the frequency response characteristics of a telephone channel. Zero mean white noise $w[n]$ with variance σ_w^2 is passed through a separate filter $\hat{G}(z)$ to produce coloured noise $v[n]$, which simulates the channel noise. The coloured noise $v[n]$ is added to the output of $\hat{H}(z)$ to obtain an approximation of telephone bandwidth speech $y[n]$.

The task is then to design $\hat{H}(z)$, the channel model, and $\hat{G}(z)$, the noise model, such that the output $y[n]$ exhibits the characteristics and degradations typical of telephone speech. In the remainder of the paper the analysis procedures followed to accomplish the above are described. It should be noted that in the application of speech recognition systems, the phase response of the filters are of little importance and therefore linear-phase filters were used for both $\hat{H}(z)$ and $\hat{G}(z)$.

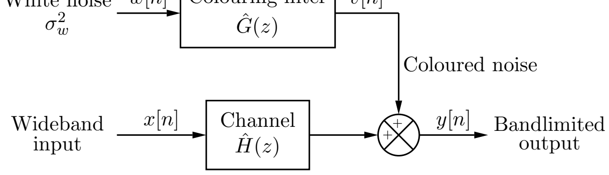

Figure 1: Model of the telephone channel.

4. Channel analysis

This section describes the techniques used to model the frequency responses of the telephone channels used to compile the NTIMIT database. Several techniques were evaluated.

4.1. Parametric channel modelling

Moving average (MA), autoregressive (AR) and autoregressive moving average (ARMA) models were considered as candidates for $\hat{H}(z)$ [2, pp. 106–118], [3, pp. 836–841], [4]. In each case, the application of the least-squares criterion allows a set of linear equations to be found for determining the model coefficients. However, in the case of the AR and ARMA models, the classical least-squares optimisation is not applied directly to the output estimation error. This lead to consistently better performance (in terms of the average output error) by MA models. It was therefore decided to focus on the evaluation of MA model topologies for the filter $\hat{H}(z)$.

In [3, pp. 882–883] the equations used to fit an MA model based on known input and output sequences $x[n]$ and $y[n]$ are presented. The least-squares criterion yields a set of linear equations

$$r_{yx}[j] = \sum_{k=0}^{q} b_k \cdot r_{xx}[j-k] \text{ for } j = 0, 1, \ldots, q \qquad (1)$$

for determining the coefficients $b_0, b_1, \ldots, b_q$ of the MA system

$$\hat{H}(z) = \sum_{k=0}^{q} b_k \cdot z^{-k} \qquad (2)$$

In equation (1), $r_{xx}[j]$ is the autocorrelation of $x[n]$ and $r_{yx}[j]$ is the crosscorrelation of $y[n]$ with $x[n]$. The crosscorrelation of $y[n]$ with $x[n]$ is defined as [3, p. 118]

$$r_{yx}[j] = \sum_{n=-\infty}^{\infty} y[n] \cdot x[n-j] \qquad (3)$$

By solving for the coefficients $b_0, b_1, \ldots, b_q$ of the MA model using equation (1), using a TIMIT utterance as input $x[n]$ and the corresponding NTIMIT utterance as output $y[n]$, an approximation of the telephone channel characteristic can be found.

4.2. Spectral channel analysis

In addition to MA, AR and ARMA models, three different spectral analysis techniques were evaluated. These techniques approximate the amplitude response of an unknown discrete-time system based on known input and output sequences $x[n]$ and $y[n]$. All three techniques are based on the short-time Fourier transform (STFT).

In all three cases, both $x[n]$ and $y[n]$ are divided into L frames of M samples as follows:

$$x_i[n] = x[n + i \cdot D] \text{ for } \begin{array}{l} n = 0, 1, \ldots, M-1 \\ i = 0, 1, \ldots, L-1 \end{array} \qquad (4)$$

In equation (4), M is the frame length and D the frame skip. Next, the DFT of each frame is taken:

$$X_i[k] = \text{DFT}\{x_i[n] \cdot w[n]\} \text{ for } i = 0, 1, \ldots, L-1 \qquad (5)$$

Here $w[n]$ is a Blackman window function. Consequently, the STFT of each sequence is determined.

For Method I, the frames of the STFT of both the input and output are averaged to obtain average magnitude spectra:

$$\left|X_{\text{avg}}[k]\right| = \frac{1}{L} \sum_{i=0}^{L-1} \left|X_i[k]\right| \qquad (6)$$

and

$$\left|Y_{\text{avg}}[k]\right| = \frac{1}{L} \sum_{i=0}^{L-1} \left|Y_i[k]\right| \qquad (7)$$

These average magnitude spectra are used to determine an approximation of the amplitude response of the unknown channel:

$$\left|\hat{H}[k]\right| = \frac{\left|Y_{\text{avg}}[k]\right|}{\left|X_{\text{avg}}[k]\right|} \qquad (8)$$

In Method II, the ratio of each of the corresponding frames from the output STFT to the input STFT is taken first:

$$H_i[k] = \frac{Y_i[k]}{X_i[k]} \qquad (9)$$

and the average magnitude determined second:

$$\left|\hat{H}[k]\right| = \frac{1}{L'} \sum_{i=0}^{L-1} \left|H_i[k]\right| \qquad (10)$$

to once again obtain an approximate amplitude response of the system.

The ratio in equation (9) has been found to be very sensitive to estimation inaccuracies when the spectral components in $X_i[k]$ are small. In such situations, distortions introduced by the windowing operation, in the form of spectral sidelobes or introduced nulls, can have a noticeable effect on the spectral estimate $H_i[k]$ and by implication $\left|\hat{H}[k]\right|$. In order to avoid this, a threshold amplitude is set for $X_i[k]$, below which terms are excluded from the average. For every value of k in equation (10), L' is taken as the number of valid points in $H_i[k]$ (i.e. the number of values above the threshold). Based on experimentation, a threshold of 15 dB below the maximum amplitude of $x[n]$ was chosen.

In Method III, the ratio of each of the corresponding frames from the output STFT to the input STFT is determined as in Method II, but instead of taking the mean in equation (10), the median of the valid values of $\left|H_i[k]\right|$ is determined. For Method III, as for Method II, a threshold value is used to exclude values of $H_i[k]$ from the median calculation.

4.3. Evaluation of channel modelling techniques

The MA model and the three spectral analysis techniques described in the preceding sections were evaluated first by using synthetic filters. Following this, evaluation using a TIMIT/NTIMIT utterance pair was performed.

The evaluation procedure using synthetic filters is illustrated in Figure 2. Three synthetic IIR bandpass filters with different cut-off frequencies were employed and each of the four

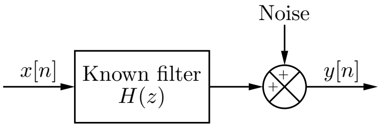

Figure 2: Evaluation of a channel model using a synthetic filter.

Table 1: Prediction errors ($\times 10^{-3}$) without noise.

Method	Filter 1	Filter 2	Filter 3	Average
MA model	0.0429	0.0425	0.0350	0.0401
Method I	2.8309	2.4953	2.0037	2.4433
Method II	2.0691	1.0720	1.0216	1.3876
Method III	2.4170	1.1865	1.0908	1.5648

Table 2: Prediction errors ($\times 10^{-3}$) with noise.

Method	Filter 1	Filter 2	Filter 3	Average
MA model	1.6456	1.7209	1.6432	1.6699
Method I	70.737	76.796	67.842	71.792
Method II	12.421	15.687	12.100	13.402
Method III	9.2653	10.714	8.2760	9.4185

techniques was evaluated by determining the spectral error:

$$E = \frac{1}{M} \sum_{k=0}^{M-1} \Big| |H[k]| - |\hat{H}[k]| \Big| \qquad (11)$$

The error defined in equation (11) is the absolute mean difference between the known amplitude response of the synthetic filter $|H[k]|$ and the approximated amplitude response $|\hat{H}[k]|$, each determined at M frequencies, with

$$H[k] = H(e^{j\omega})\big|_{\omega=2\pi k/M} \text{ for } k = 0, 1, \ldots, M-1 \qquad (12)$$

An arbitrary TIMIT utterance, with a variance of $580 \cdot 10^{-6}$, was used as input $x[n]$. The prediction errors were determined both using a clean input $x[n]$ and when introducing additive Gaussian white noise with a variance of $\sigma_w^2 = 100 \cdot 10^{-9}$ as indicated in Figure 2. The results for the clean speech are shown in Table 1 while, for the noisy speech, average prediction errors for 100 repetitions of the experiment are shown in Table 2. The results indicate that the MA model yields the smallest prediction error and Method I the worst performance in both cases. When no noise is added (Table 1), Method II is the spectral analysis technique with the best performance. However, when noise is added (Table 2), Method III shows the lowest prediction error among the spectral analysis techniques.

Based on these results, the MA model and Method III were chosen for further evaluation. A single TIMIT/NTIMIT utterance pair was used to approximate the amplitude response of a real telephone channel. The approximate amplitude responses, according to the MA model as well as Method III, are presented in Figure 3 and Figure 4 respectively.

Figure 3 shows significant attenuation for frequencies higher than 3.2 kHz, which would be expected for typical telephone channels [5]. However, significant attenuation would also be expected for frequencies lower than 300 Hz, but this is not reflected in the response shown in Figure 3. Further analysis showed this anomaly to be caused by the presence of low frequency noise, discussed in more detail in Section 5.

Figure 4 shows significant attenuation for frequencies lower than 150 Hz and higher than 3.2 kHz. In general it was observed that Method III shows more resilience to the effects caused by the low frequency noise. The amplitude threshold set for $X_i[k]$, discussed in Section 4.2, is used to exclude terms from the median when spectral components of the input frame are small. Because the energy of the unwanted low frequency noise is always much lower than that of the speech, it dominates only in output frames with a low signal-to-noise ratio (i.e. in silence frames). When the signal-to-noise ratio of an output frame is low, the spectral components of the input frame will generally be small. Therefore, the amplitude threshold effectively eliminates spectral components from such frames from the channel estimation, thus improving robustness. This leads to the conclusion that Method III will be less influenced by the presence of the low frequency noise.

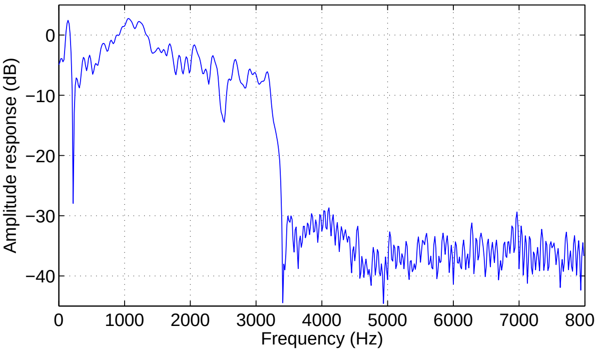

Figure 3: Approximate amplitude response of the NTIMIT channel using the MA model.

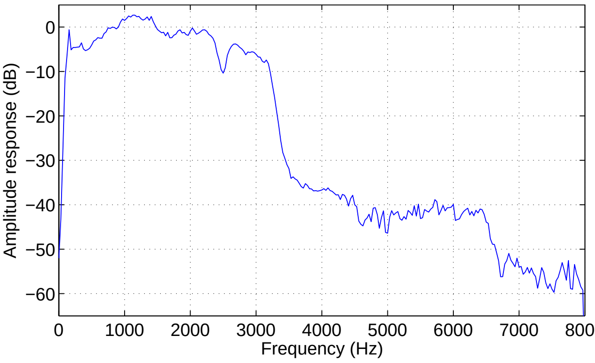

Figure 4: Approximate amplitude response of the NTIMIT channel using Method III.

By comparing Figure 3 and Figure 4, it is also clear that the latter shows more detailed response characteristics for higher frequencies. For example, in Figure 4 the effect of the anti-aliasing filter, with a cut-off frequency of 6.4 kHz, can clearly be seen while these details are absent from Figure 3. Based on these observations, Method III was selected as the technique for the analysis of the frequency response characteristics of a telephone channel.

4.4. Analysis using TIMIT/NTIMIT utterances

Using Method III from the preceding section, the frequency response characteristics of the telephone channels used to create the NTIMIT database were analysed in an attempt to determine the channel model $\hat{H}(z)$ in Figure 1. Approximate amplitude responses were obtained for each of the 6 300 TIMIT/NTIMIT utterance pairs and grouped according to the central office to which each TIMIT utterance was transmitted. Subsequently 253 groupings at $M = 512$ frequencies were obtained.

For each group the mean amplitude response was determined. Two of these responses were unusable due to erroneous files in the NTIMIT database, leaving 251. Each of these responses can be seen as a legitimate possible response for the filter $\hat{H}(z)$ in Figure 1. This approach, however, limits the channel filter to only 251 possibilities.

To address the limited number of filter responses that can be obtained in this way, further analysis was performed. At each frequency the mean and the variance of the 251 amplitude responses, expressed in dB, were determined. The mean amplitude response of the NTIMIT channels determined in this way and the interval given by its standard deviation are presented in Figure 5. A histogram of the 251 values at $f = 1.25$ kHz and the associated estimated Gaussian probability density function are illustrated in Figure 6.

Figure 5 shows significant attenuation for frequencies lower than 300 Hz and for frequencies higher than 3.2 kHz and a sharp notch at about 2.5 kHz. Within the telephone network, tones are often used for communication between telephone offices. The frequency of these tones is specified as 2.6 kHz in [5], although Figure 4 suggests a slightly lower frequency. The effect of the anti-aliasing filter is also apparent.

Figure 6 suggests that a Gaussian probability density function provides an acceptable estimate of the distribution of the values of the amplitude responses at $f = 1.25$ kHz. Values at other frequencies exhibited similar characteristics. By modelling the response of the telephone channel as being Gaussian-distributed about the mean response shown in Figure 5, an infinite number of channel characteristics can be generated. This liberates the simulation of the channel characteristic from the 251 "prototype" filter characteristics that can be extracted from TIMIT/NTIMIT.

5. Noise analysis

In order to analyse the additive noise of the NTIMIT channels, segments of audio where no speech is present were isolated. A total of 100 segments from 100 arbitrarily chosen NTIMIT utterances were extracted for the noise analysis. The assumption was made that these noise segments would give an acceptable indication of the characteristics of the noise which is added to the speech signal by the telephone channel.

5.1. The Yule-Walker equations

Each noise segment was assumed to correspond to a coloured noise sequence $v[n]$ as indicated in Figure 1. The problem is then to obtain the filter $\hat{G}(z)$ for a given $v[n]$. The Yule-Walker equations allow for the derivation of the parameters of an AR model which is excited by a zero mean white noise sequence with variance σ_w^2 when the output of the system is known [2, pp. 114–118]. This agrees precisely with the noise model presented in Figure 1.

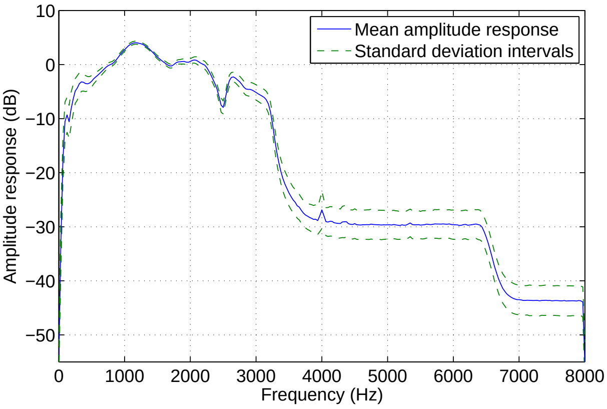

Figure 5: Mean amplitude response of the NTIMIT channels with the standard deviation intervals.

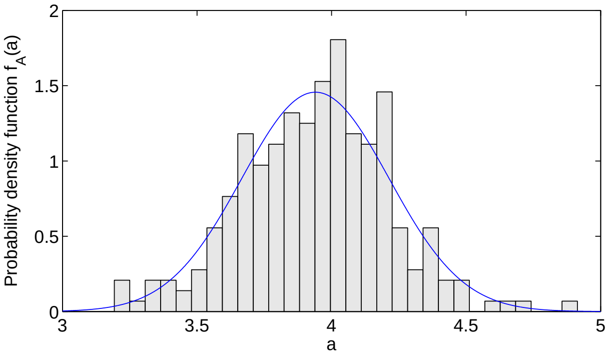

Figure 6: Histogram of the amplitude responses of the NTIMIT channels at $f = 1.25$ kHz and the superimposed Gaussian approximation.

The unknown coefficients of the AR model in equation (13) are determined using the Yule-Walker equations given in equation (14) and the variance of the white noise sequence can be determined using equation (15) once the coefficients $a_1, a_2, \ldots, a_p$ are known.

$$\hat{G}(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k \cdot z^{-k}} \tag{13}$$

$$r_{vv}[j] = -\sum_{k=1}^{p} a_k \cdot r_{vv}[j-k] \text{ for } j = 1, 2, \ldots, p \tag{14}$$

$$r_{vv}[0] = -\sum_{k=1}^{p} a_k \cdot r_{vv}[-k] + N \cdot \sigma_w^2 \tag{15}$$

The Yule-Walker equations can also be shown to agree exactly with the equations used to determine the coefficients of a linear prediction (LP) filter [2, pp. 156–158]. In the remainder of the paper, filters obtained using the above equations are therefore referred to as LP filters.

5.2. Analysis of the noise segments

Using the equations given in the preceding section, LP filters were obtained for each of the 100 noise segments. Figure 7 shows the average and the median frequency response for the 100 filters, along with the 90% intervals.

Figure 7 shows approximate harmonics of 120 Hz at 121 Hz, 234 Hz and 359 Hz. An audible hum was present in some of the noise segments and these harmonics are possibly an indication of the presence of a periodic component in the noise. A frequency of 60 Hz is used for power distribution in the USA, where NTIMIT was compiled, which could possibly account for these components. It is also stated in [5] that prominent signal energy in the area between 40 Hz and 60 Hz was detected during the compilation of the NTIMIT database. This was not seen as a typical characteristic of the telephone channel and hence the entire set of NTIMIT utterances was filtered to remove these unwanted components. It is possible that the low frequency peaks in the average LP spectrum are in fact the remaining harmonics

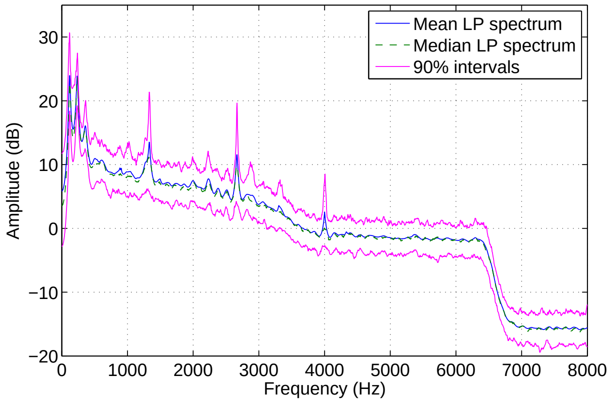

Figure 7: Mean and median LP spectra with the 90% intervals from the median.

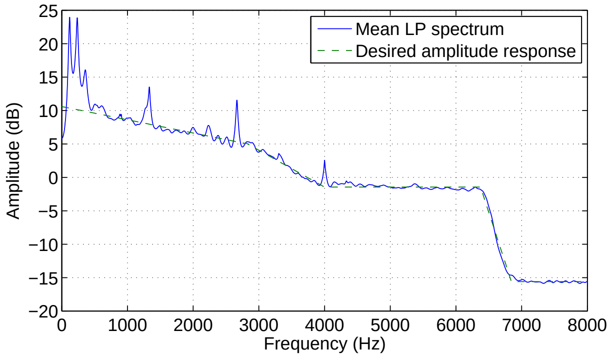

Figure 8: Mean LP-spectrum and desired amplitude response for the FIR colouring filter $\hat{G}(z)$.

of this filtered noise. Harmonics at 1.3 kHz, 2.6 kHz and 4 kHz can also be distinguished (harmonics of 1.3 kHz).

Whether these harmonics are the result of the power network (which would mean that the frequency would differ according to the national power distribution frequency) or whether they are due to an anomaly in the creation of the NTIMIT database is not known. It was however assumed that these periodic components are not part of the characteristics of a typical telephone channel and hence they were not considered in the design of the colouring filter $\hat{G}(z)$.

The amplitude response of $\hat{G}(z)$ was designed based on the mean LP spectrum shown in Figure 7. A piecewise linear function was used to specify the desired amplitude response at 257 frequencies from $f = 0$ Hz to $f = 8$ kHz, which resulted in 512 points across the spectrum. The desired amplitude response is shown in Figure 8. Using the windowing technique, a linear-phase FIR filter with the desired amplitude response was obtained [3, pp. 664–670]. Thus a FIR colouring filter for $\hat{G}(z)$, as shown in Figure 1, was determined.

6. Experimental evaluation

This section considers the application of the techniques described in the preceding sections for the simulation of telephone channels given clean speech. First, a single TIMIT/NTIMIT utterance pair is considered. Next, the techniques are evaluated by application to the development of ASR systems.

6.1. Obtaining the channel and noise models

As stated in Section 4.4, a library of 251 amplitude responses representative of typical telephone channels, which each can be used as the channel filter $\hat{H}(z)$, was extracted from the TIMIT and NTIMIT databases. This gives one method of selecting the channel filter. As an alternative, the Gaussian distributions determined at the 512 frequencies can be used to generate a random filter characteristic (described in Section 4.4). The advantage of this technique is that an unlimited number of different channel models can be obtained.

However, this second approach assumes the amplitude response at adjacent discrete frequencies to be statistically independent, which is untrue. A randomly generated filter thus has a jagged amplitude response which is undesirable. This was addressed by applying a narrow window to the filter impulse response to achieve smoothing in the frequency domain. The amplitude response of a filter generated in this way is shown in Figure 9.

The design of the noise colouring filter $\hat{G}(z)$ was described in Section 5.2. Techniques for determining both the channel and noise models were thus found and consequently the model in Figure 1 can be used for the simulation of telephone channels.

6.2. Application to single NTIMIT channel

Using the Welch method, the power density spectrum (PDS) of a single TIMIT utterance and the PDS of the corresponding NTIMIT utterance were estimated and are shown in Figure 10 [3, pp. 975–977]. The bandlimiting effect of the telephone channel is apparent.

The same TIMIT utterance was used as input $x[n]$ in Figure 1 and $\hat{H}(z)$ was chosen from the channel library to correspond to the channel used to record the specific NTIMIT utterance. The noise colouring filter $\hat{G}(z)$ was designed as described in Section 5.2 and $w[n]$ was taken as zero mean Gaussian white noise with a variance of $\sigma_w^2 = 26.5 \cdot 10^{-9}$.

The estimated PDS of the approximated telephone-quality speech $y[n]$ is presented in Figure 11, together with the estimated PDS of the NTIMIT utterance. The PDS of $y[n]$ is a close approximation of the PDS of the NTIMIT utterance. To determine the effect of the noise model, the variance of the noise $w[n]$ was reduced to $\sigma_w^2 = 0$. The estimate of the PDS of the output $y[n]$ obtained in this manner is also shown in Figure 11. The contribution of the noise model is apparent from the comparison between the PDS of $y[n]$ with noise and the PDS of $y[n]$ without noise.

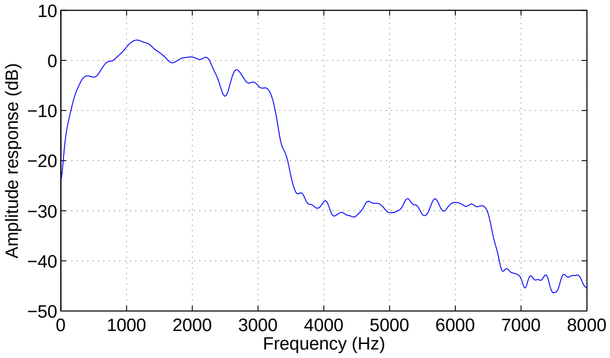

Figure 9: Amplitude response of a generated filter.

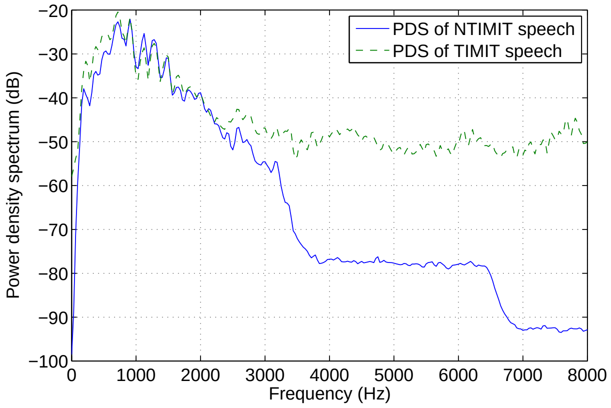

Figure 10: Estimated PDS of the TIMIT and NTIMIT utterances.

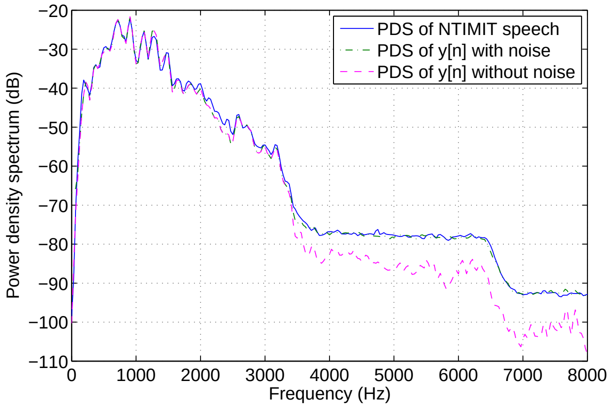

Figure 11: Estimated PDS of the NTIMIT utterance and the approximate telephone-quality speech, with and without noise.

6.3. Evaluation in the development of ASR systems

Using HTK [6], ASR systems were developed based on a variety of training sets. Monophones were trained using 3 state left-to-right HMMs with 8 mixture diagonal covariance Gaussian distributions for each state. 13 MFCCs with the first and second derivatives were used to obtain 39 dimensional feature vectors. The underlying filterbank analysis used 18 channels spanning 300 Hz to 3.4 kHz as a first approximation of a telephone channel. Phone recognition experiments using a phone loop grammar were performed for performance evaluation.

Four systems were developed by applying the same training process to four different training sets. Each system was evaluated by means of the standard NTIMIT test set. The first system was trained on the standard NTIMIT training set and serves as a baseline with matched test/train conditions. The second system was trained on the 300 Hz to 3.4 kHz bandpass filtered TIMIT data and represents a crude but often-used approximation of telephone speech.

The third and fourth systems were trained on TIMIT utterances that had been filtered using generated filter characteristics for $\hat{H}(z)$ in Figure 1, as described in Section 6.1. The third system included noise with an SNR of 30 dB using the noise colouring filter $\hat{G}(z)$ described in Section 5.2, while the fourth system included no noise. The results of the ASR experiments are shown in Table 3.

Table 3: Accuracies for the experiments using ASR systems.

Training set	Test Set	% Accuracy
NTIMIT	NTIMIT	40.65%
TIMIT narrowband	NTIMIT	32.56%
Filtered TIMIT, 30 dB noise	NTIMIT	36.34%
Filtered TIMIT, no noise	NTIMIT	32.19%

Although the accuracy obtained using the third system is 10.6% lower than the accuracy using the NTIMIT training set, it results in an 11.6% increase in accuracy from the basic bandpass approach to training. However, when no noise is added, performance is not much different from the TIMIT approach. This leads us to conclude that, from an ASR perspective, the noise model is by far the most important aspect of the complete model. The reasons for this are the subject of ongoing work.

7. Summary and conclusion

The development of techniques that allow the effects of a variety of telephone channels to be simulated has been presented. Different techniques for the approximation of frequency responses from the TIMIT/NTIMIT utterances were evaluated and a method based on the median of frame-wise spectral estimates was found to be most robust to the noise levels present in these corpora. Two techniques for the design of a channel model were proposed: the first based on a channel library and the second based on the generation of filters according to estimated distributions of the amplitude responses. A noise model was developed based on the analysis of silence segments from the NTIMIT utterances. LP filters driven by Gaussian white noise were found to give a good spectral approximation to the measured noise signals. The techniques were shown to closely mimic the spectral changes brought about by the telephone channels in the NTIMIT database. The application of the techniques in the development of ASR systems indicates that the noise model is the major contributing factor to an increase in accuracy relative to a bandpass channel approximation.

8. References

[1] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, “NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Silver Spring, MD, 1990, pp. 109–112.

[2] S.M. Kay, *Modern Spectral Estimation: Theory and application*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[3] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, algorithms, and applications*, 4th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2007.

[4] C.L. Phillips and H.T. Nagle, *Digital Control System Analysis and Design*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1995, pp. 406–410.

[5] C. Jankowski, *The NTIMIT Speech Database*, Printed documentation which accompanies the NTIMIT CD-ROM, January 1991.

[6] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, 2002.