ESTIMATION OF FORMANTS IN NOISE CORRUPTED SPEECH USING AUDITORY MODELS.

F. K. Fink and P. Dalsgaard

Speech Technology Centre, Institute of Electronic Systems,
University of Aalborg, Fredrik Bajers Vej 7,
DK-9220 Aalborg, DENMARK.

ABSTRACT

In many practical applications involving speech recognition it is of great importance to be able to handle noise suppression in the preprocessing stage. In this paper we describe different front-end processing systems, one based on speech production modelling and two based on auditory modelling, and present results on their formant estimation abilities when being excited by speech signals contaminated by noise. Three noise types - car, cocktail-party and open-plan office noise - are added to speech signals at signal-to-noise ratios varying between 20 and -10 dB. The results show that preprocessing using auditory modelling is much more robust to noise than speech production modelling, and that formants can still be reliably estimated at SNR = -5 dB for speech signal contaminated by car noise.

1. INTRODUCTION

Work within estimation of acoustic-phonetic parameters, e.g. formants, from speech signals has so far to a large extent been based on signal analysis employing speech production modelling and using speech signals containing noise or only a small amount of noise. For many practical applications such as those in connexion with automatic speech recognition (ASR) or devices used to assist profoundly deaf people in lip reading [1], it is of great importance to maintain the ability to estimate these parameters robustly even from speech signals contaminated by environmental noise.

Research in auditory modelling and signal analysis using these models have shown evidence, that the non-linear processing in the auditory models to a certain degree is able to suppress noise, such that essential acoustic-phonetic information can still be estimated from the speech signal. The aim of this paper is to report work in this field using two different approaches to auditory modelling, and demonstrate the quality of these models in respect of noise suppressing properties are measured against a standard speech production model.

The acoustic-phonetic parameters analysed here are the formants, and the data used in the work are taken from the EUROM.0-database, which has been established within the ESPRIT project SAM (Speech Assessment Methodology).

Modelling of the auditory reception of speech can be carried out basically from two major approaches.

Auditory models based on psychoacoustic measurements [2] are frequency domain models. Here the frequency selectivity of the basilar membrane is simulated on the basis of Fourier Transformation of the speech signal. The psychoacoustic model furthermore consists of simulation of psychoacoustic characteristics such as the equal loudness curve, Stevens power law and intensity to loudness conversion.

Models based on neurophysiological measurements [3],[4] are time domain models initially containing a critical parallel filterbank followed by two nonlinear stages. The first nonlinear stage is modelling the most prominent time domain features of the transformation from the vibration of the basilar membrane to the firing rate in the auditory nerve. The last stage of the model used in the work presented here is the synchrony detector, which measures the extent to which the signal in each channel is dominated by information near the center frequency of the specific critical band.

The most characteristic difference between the psychoacoustic and the neurophysiological model are the different nonlinear stages and the synchrony detection. In agreement with results presented by Ghitza [5], we found that the linear filtering part of both models is not critical.

Recently models containing elements from both psychoacoustic and neurophysiological models have shown promising results in estimation of parameters from speech signals contaminated by noise [6].

The purpose of this paper is to compare two fundamentally different signal analysis methods, namely 1) speech perception modelling and 2) speech production modelling in respect of their ability to estimate formants from speech signals being subjected to varying levels of noise and noise types. The additive noise signals used are

1) car noise collected in a driving automobile,
2) cocktail-party noise and
3) open-plan office noise,

and the signal-to-noise ratios of the simulated speech signals varies between 20 dB and -10 dB.

The following sections describe some details of the SAM-database, the signal-to-noise calculations, the speech production model, the auditory models and the results.
2. DATABASE

The continuous speech data used in this work is taken from the SAM-database [7]. This database contains a part with isolated words and connected words and four passages of each two minutes duration spoken by two females and two male speakers. The SAM-database is stored on a CD-ROM and is thus easily accessible from a Personal Computer. The database contains similar speech material from five European languages, and all spoken passages have been manually segmented and labelled by expert phoneticians to serve as reference material.

The sampling rate of the SAM-database is 16 kHz. The data used within this work is downsampled and bandwidth limited to telephone quality.

3. NOISE TYPES

In order to serve as reference noise for the analysis performed here we have collected three types of noise, which is believed to cover a realistic range of practical user environments. These are car noise, cocktail-party noise and open-plan office noise. The car noise is collected in a driving automobile under different driving speeds, road conditions and with windows open or closed. Similar varying external conditions are taken into consideration during data collection with the other two types of noise in order to ensure the optimal environmental simulations.

Spectral analysis of the noise database have shown that the car noise is dominated by rather low frequency components, whereas the cocktail-party noise is almost evenly distributed over the entire frequency range of speech signals. The cocktail-party noise and the car noise is rather homogeneous over time contrasting noise collected in an open-plan office. This contains impulsive noise sounds such as those from telephone bells and door slams.

The noise signals used in this work are added linearly to the speech data. Dependent on the signal-to-noise ratio (SNR) to be used with the different simulations reported in section 6, the sampled noise signal amplitude N(n) is being varied by a scaling factor:

\[ y = 10^{(SNR0 - SNR)/20} \]  (1)

where SNR is the signal-to-noise ratio to be used with the simulations, and SNR0 is computed by the following expression:

\[ SNR0 = 10 \log \frac{\sum_m S^2(n)}{\sum_m N^2(n)} \]  (2)

S(n) and N(n) is the sampled amplitude of the speech and the noise signal at time n, and m is the total number of samples used in the simulations. In this work m = 12372 samples (1.55 sec).

4. SPEECH PRODUCTION MODELLING

The referencing acoustic preprocessing of the speech signal is carried out using a speech production model. In this the signal is multiplied by a Hamming window and eight linear prediction coefficients are computed in windows of 16 msec duration shifted in 8 msec frames. Using a root-solving algorithm the linear prediction polynomial is split into second order sections from which the formants are identified.

5. AUDITORY MODELS

The psychoacoustic based auditory model used is developed by H. Hermansky [2]. The model is developed for the purpose of formant estimation and the method is called Perceptually based Linear Prediction (PLP).

The signal analysis carried out by this frequency domain model contains the following steps: Simulation of critical bandpass filtering based on Fourier Transformation of the speech signal, followed by a simulation of the equal loudness function, Stevens power law and intensity to loudness conversion. The output from the 18 filters of the model is interpolated linearly on a bark scale. This generates data which are used as input for AR-modelling (as for speech production modelling).

In order to adjust this model to telephone bandwidth speech signals used in this paper, some parameters are changed. The number of filters have been reduced to 14, and the size of the frequency transformation performed is reduced from 256 to 128 point DFT. This modified model has been found to give the best noise robust formant estimation using an 8th order PLP-model. New outputs are estimated every 8 msec.

The neurophysiological based model used is developed by S. Seneff and described in [3][4]. The model has been used as a preprocessor for multi-level segmentation using a sampling rate of 16 kHz.

Also this model is modified in order to handle speech signals sampled by 8 kHz and bandlimited to telephone quality. The modifications include the following: The number of filters in the filterbank is reduced from 40 to 31 in order to accomodate the bandwidth reduction. The coefficients in the remaining filters are changed in order to maintain the filter bandwidth, distance between center frequencies and overall frequency characteristic. The frequency dependent parts of the nonlinear second stage of the model are adjusted. Finally in the third stage, the general synchrony detector is resimulated to find the proper parameter values.

6. RESULTS

In this section the results from the simulations are presented. The simulation results cover two models for preprocessing, namely the speech production model and Hermansky's modified auditory model. Both models are stimulated by all three noise types under varying signal-to-noise ratios.

Alle three noise types described in section 3 are used and the simulation results gives first, second and third formant values as functions of time measured in seconds. The sampled noise signal amplitudes are scaled by the factor y (1) before added to the speech signal. SNR-values of 20, 10, 0 and -10 dB are used as indicated on the figures. Only for-
mant values in voiced speech segments are drawn for convenience.

Many simulations have to be performed in order to have sufficient confidence in the results. Due to limited space, therefore, only few graphs will be presented here showing the prominent tendencies.

Figure 1 shows the original noise free speech waveform and the corresponding spectrogram as computed by the SPIRE [8] program.

**Noise free signals.**

Figure 2a shows the three formants estimated by the speech production model. Figure 2b shows the corresponding PLP estimation of the formants. Figures 2a and 2b both show formant tracking in agreement with the tracks seen in the SPIRE spectrogram.

**SNR = 20dB.**

Figure 3 shows the formants estimated by the speech production model with car noise added to the speech signal. Comparing with Fig. 2a it is seen that the model breaks down, especially in segments where the first and second formants are close or in segments with second formants of low energy. This limited formant tracking ability has been found for all three types of noise for the given high energy signal-to-noise ratio.

Exiting the auditory model by the same signal has shown only minor changes in formant values for all three types of noise.

**SNR = 10dB.**

At this noise level the first visible deviation from the corresponding reference formants (Fig. 2b) using the auditory model has shown up. From Figure 4 it is seen, that between 0.1 and 0.3 sec. and between 0.7 and 1.0 sec. the second and third formants have almost vanished. This is not the case in the remaining segments. It is expected that this must be a result of large variations in the noise signal.

Simulation with car and cocktail-party noise have not shown this degradation for the auditory model.

**SNR = 0dB.**

Figure 5 shows results from analysis using cocktail-party noise. The speech production model (Figure 5a) now tracks the formants in an almost random way. Using the auditory model (Figure 5b) it is still possible to follow the first formant relative reliably in contrast to the second and third formant. Using open-plan office noise gives the same results.

Using car noise the auditory model still gives good estimates of the formant values with this signal-to-noise ratio.

**SNR = -5dB.**

From Figure 6 it is seen, that the auditory model still rather robustly tracks the first and second formants at this very low signal-to-noise ratio using car noise. Increasing the noise level to -10 dB signal-to-noise ratio shows that at this very high noise level the auditory model is not able to estimate the formants.

7. CONCLUSIONS

The work described in this paper demonstrates that auditory modelling is superior to speech production modelling in noise robust formant estimation. It has been found that even low noise levels influences the reliability of the estimated formants using speech production modelling. As expected the different noise types show different effect on the results from the auditory model. For all three noise types the auditory models are superior to the speech production model in respect of formant estimation capability. Measured in terms of noise level we have found this to be at least 20 dB.

Although no results from the simulations using Seneff's modified auditory model are presented here we have been using this model together with Hermansky's model in another research work also presented at this conference [9]. In this alternative work Hermansky's model and Seneff's model are used in segmentation of speech signal, and it is here shown that the performance of the two models are very close to each other.

8. REFERENCES


Figure 1: SPIRE-spectrogram of the speech signal.

Figure 2a: Formants estimated by speech production model on noise free speech signal.

Figure 2b: Formants estimated by auditory model on noise free speech signal.

Figure 3: Formants estimated by speech production model on speech signal with car noise added. SNR = 20dB.

Figure 4: Formants estimated by auditory model on speech signal with open-plan office noise. SNR = 10dB.

Figure 5a: Formants estimated by speech production model with cocktail-party noise added. SNR = 0dB.

Figure 5b: Formants estimated by auditory model with cocktail-party noise added. SNR = 0dB.

Figure 6: Formants estimated by auditory model from speech signal with car noise added. SNR = -5dB.