DETECTION OF STOP CONSONANTS IN CONTINUOUS NOISY SPEECH BASED ON AN EXTRAPOLATION TECHNIQUE

Rajyalakshmi Dokku and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
{rajyalakshmi.dokku, rainer.martin}@rub.de

ABSTRACT

In this contribution, we present an algorithm for stop consonant detection in continuous noisy speech at low signal-to-noise ratios (SNR) in the range of -5 to 7 dB. In our approach a signal extrapolation technique is used to predict the future samples based on present observations and then the predicted signal is used in the detection process. Experiments are performed over 700 utterances of 60 speakers taken from the TIMIT database with three different noise types. The detection performance achieved is 71% at -5 dB input SNR and 90.21% at 7 dB input SNR in case of white Gaussian noise at false alarm rates of 25.14% and 10.81% respectively. The detection performance results achieved are well in line with previous methods though our experiments are conducted at lower SNRs.

Index Terms— Signal detection, Extrapolation, Autoregressive process.

1. INTRODUCTION

One of the main problem faced by hearing impaired people is speech recognition in noise. While vowels are relatively easy to detect, the detection of stop consonants in noisy environments is much more difficult because of their abnormal loudness. Stop consonants (/b/, /d/, /g/, /k/, /p/, and /t/) are transient sounds comprising a short pause followed by a short impulse-like burst. As shown in [1], stop consonants contain rich information and are important for speech recognition for normal and hearing impaired listeners. So, they deserve special treatment during noise reduction and speech compression compared to the accompanying vowels.

At low input SNRs, voice activity detection (VAD) based noise reduction algorithms often consider stop consonants as noise because of their structure and their presentation levels much below those of vowels in speech signals. As a results the overall intelligibility of the enhanced signal is reduced. In [2] the effect of preservation of stop consonants on a noise reduction system was investigated. It showed that recognition of stops significantly improved when the release burst of the stop was amplified.

Ali et al. [3] conducted experiments for stop consonant place detection and classification in the range of 5 to 60 dB input SNRs. The detection of stop consonants in the presence of noise at input SNR below 0 dB is a most challenging task. In this paper we implement an algorithm for the detection of stop consonant frames in continuous noisy speech based on an extrapolation technique. This algorithm can be used as a pre-processing block to a noise reduction or a speech recognition system. In our experiments, we use four speech features. They are the prediction gain and energy difference derived based on an extrapolated signal, a periodicity feature and the zero-crossing rate.

The short duration of the burst makes stop consonants vulnerable to model by an autoregressive (AR) process. Examining the AR process performance during stop consonants gives us an important intuitive guidelines for the classification between stationary and non-stationary components. We also show the algorithm's performance using the measure of correctness taking manually labeled speech as a reference. The statistics of the detector was taken over 700 speech files mixed with different types of noise.

The remainder of this paper is organized as follows. In Sec. 2 we describe the extrapolation technique along with AR model order selection criteria. The implementation of the stop consonant detection algorithm is presented in Sec. 3. In Sec. 4, we demonstrate the detector performance results for three noise types. We conclude our work in Sec. 5.

2. EXTRAPOLATION TECHNIQUE

The extrapolation is a method to extend speech samples in forward or in backward direction based on present observations. In our context extrapolation of a finite length signal vector $X = [x_1, x_2, x_3, ..., x_N]$ means the calculation of new future unknown samples $X_{Extr} = [x_{N+1}, x_{N+2}, x_{N+3}, ...]$. In this technique we assume that there exists a set of prediction filter coefficients $a = [a_1, a_2, ...a_p]$ of order p that would linearly predict any sample in a given signal perfectly with zero prediction error [4], i.e.,

This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} \qquad (1)$$

If we have at least p known samples in the given signal vector we can generate the first forward extrapolated sample x_{N+1} by the above equation resulting in a prolonged signal vector $X = [x_1, x_2, x_3, ..., x_N, x_{N+1}]$. Now the last p samples are used to predict the second forward extrapolated sample x_{N+2} using above equation again. By successively using this procedure we can generate new samples.

There are several ways to obtain the prediction coefficients of the signal, e.g. by solving the autoregressive (AR) model parameters. In this paper we computed AR parameters for an extrapolation.

2.1. AR Process Theoretical Background

The discrete noisy signal model is given by,

$$y_k = x_k + n_k \quad \forall\, k \qquad (2)$$

where y_k is observed signal, x_k is the clean speech signal, n_k is noise signal and the suffix k represents the sample index.

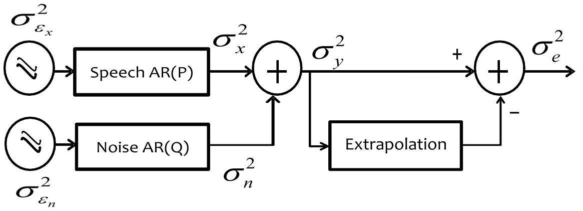

Fig. 1. Block diagram for error variance computation in between the processed noisy speech and the extrapolated noisy speech.

As shown in Fig.1, x_k and n_k are assumed to be generated from two independent AR processes of orders P, Q respectively.

$$x_k + a_1 x_{k-1} + a_2 x_{k-2} + ... + a_P x_{k-P} \equiv \epsilon_x \qquad (3)$$

$$n_k + b_1 n_{k-1} + b_2 n_{k-2} + ... + b_Q n_{k-Q} \equiv \epsilon_n \qquad (4)$$

where a_j, b_j are model coefficients and ϵ_x,ϵ_n are white stationary random processes with zero mean and variances $\sigma^2_{\epsilon_x}$, $\sigma^2_{\epsilon_n}$ respectively.

According to equation (2), the noisy signal model y_k is rewritten as:

$$y_k = -\sum_{j=1}^{P} a_j x_{k-j} + \epsilon_x - \sum_{j=1}^{Q} b_j n_{k-j} + \epsilon_n \qquad (5)$$

As shown in [5], the sum of two independent AR processes results in an ARMA process, i.e.,

$AR(P) + AR(Q) = ARMA(P + Q, \; max(P, Q))$
An $ARMA(P + Q, \; max(P, Q)$ process can always be approximated by a higher order MA model and in [5], it is also shown that an $MA(Q)$ process can also be approximated by an higher order AR process. In practice, the ARMA model is more difficult to fit to data than the AR model. So, in this paper we used a high order AR model to approximate the noisy speech signal.

For computing the AR model parameters, there are several estimation methods available, the most prominent of which are maximum likelihood (ML) assuming Gaussian errors, least squares (LS), approximate maximum likelihood (aML), YuleWalker (YW) and the Burg method [6].

Among the above mentioned estimation methods in this work we use Burg's method for computing AR parameters. The major advantages of the Burg method for estimating the parameters of the AR model are its high frequency resolution, it yields a stable AR model and it minimizes the both forward and backward prediction errors while computing the AR coefficients. Detailed derivation of Burg's method along with the Levinson-Durbin recursion found in [6].

By substituting the above computed AR parameters in equation (1), we can compute the extrapolated samples, but the question remains which model order should be used. To investigate which model order is suitable for modeling the input signal using extrapolation, we considered AR model order selection criteria in the following section.

2.2. AR Model Order Selection

To model the time series with an AR process, we first need to determine the model order of the process. Automatic order selection using statistical order-selection criteria was first introduced by Akaike and then, many other modeling techniques have been evaluated. Those are Final Prediction Error (FPE), Akaike Information Criteria (AIC), Kullback Information Criteria (KIC), Residual Variance (RV) and Minimum Description length (MDL) [6].

Model order selection purely depends on the type of input data given to the model. For instance, a signal which shows rapid variations and transients will require a relatively large model order. The mean-square value of the residual error is one of the good parameters to measure the performance of the AR model.

In this paper we implemented a better known criteria, AIC on extrapolated data for selecting the model order along with the mean square error (MSE). AIC is a measure of goodness of fit of an estimated statistic model. AIC reflects the balance between complexity of model order and goodness of fit and is given by,

$$\text{AIC}(p) = \ln(\sigma_e^2) + \frac{2p}{N} \qquad (6)$$

where N is the number of data points used and σ_e^2 is the residual variance and p is the model order. The term $\frac{2p}{N}$ in equation (6) gives the penalty for the selection of higher orders.

The simulation results of AIC and MSE for different filter orders at different input SNR levels are computed on the extrapolated data and which are shown in Fig. 2. In Fig. $2(a)$, MSE curves shows minimum at the filter orders in the range of $60-80$ and then its starts increasing again. In Fig. $2(b)$, the curves of AIC at different input SNR levels also shows minimum values and slow variations around the filter order 60, means where the penalty factors increasing faster for every increase of filter order. As a conclusion, the filter orders in the range of 60-80 are preferable for this application (extrapolation).

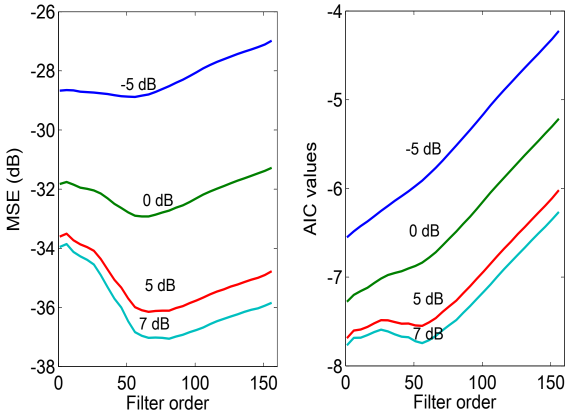

Fig. 2. Shows the simulation results of MSE and AIC versus filter orders at different input SNR levels.

3. ALGORITHM

The algorithm as shown in Fig. 3 is divided into two parts. In the first part (Step 1) the classification is done in between stationary and non-stationary frames by using statistical properties of signal modeling. During the second part (Steps 2 and 3) stop consonants are extracted from non-stationary components by using physiological features (pitch, ZCR) of speech production.

The detection algorithm works frame-by-frame. In this process at first the noisy speech signal y_n is segmented into M frames with the frame length N and frame advance R. Then, the i^{th} noisy frame is represented as $Y_i = [y_{iR+1}, y_{iR+2}, y_{iR+3}, ..., y_{iR+N}]^T$. In Fig. 3, I is the vector of the size $1 \times M$, storing the frame index values.

After the segmentation of the noisy speech, by giving $(i-1)^{th}$ frame Y_{i-1} as input to the extrapolation block, N forward extrapolated samples are computed by using the extrapolation technique as explained in Section 2. All newly computed N extrapolated samples together form an extrapolated frame which is represented as Y_{extr_i}. The first frame of the noisy signal and the extrapolated signal are the same because no previous information is available to predict the first frame. All other frames of the extrapolated signal are calculated from their previous frames of the noisy signal. The amount of predictability is measured by comparing the extrapolated frame and corresponding original noisy frame in terms of prediction gain and energy difference.

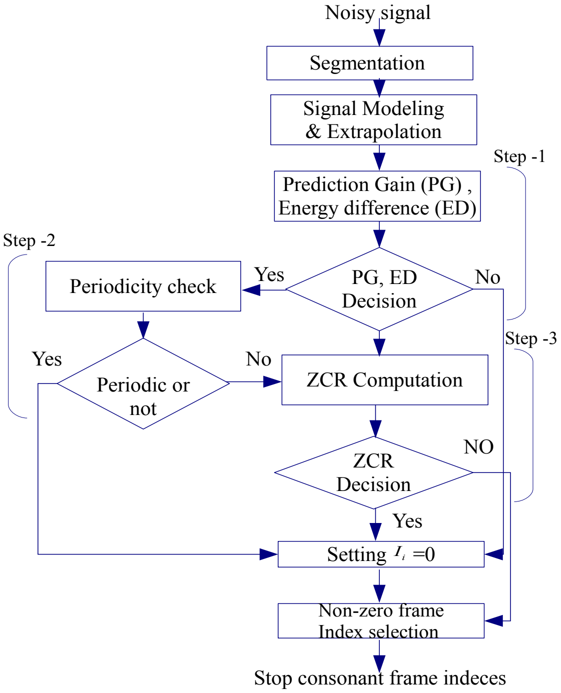

Fig. 3. Stop consonants detection algorithm flowchart.

3.1. *Step 1:* Prediction Gain (PG)

The prediction gain and energy difference are the parameters used in this work for measuring the success of the prediction. The PG and the energy difference (ED) are computed as shown in equations (7) and (9), respectively,

$$PG_i = 10\log_{10}\left(\frac{\sigma_y^2}{\sigma_e^2}\right) = 10\log_{10}\left(\frac{\sigma_x^2+\sigma_n^2}{\sigma_e^2}\right) \quad (7)$$

$$EO_i = Y_i^T Y_i \ ; \ EE_i = Y_{extr_i}^T Y_{extr_i} \quad (8)$$

$$ED_i = |10\log_{10}(EO_i) - 10\log_{10}(EE_i)| \quad (9)$$

where ED_i is the energy difference between the noisy and the extrapolated signal of frame i. The PG decreases with decreasing input SNR because the noise is more dominant. The residual variance σ_e^2 decreases with increasing AR model order. For white noise, during low input SNR's $\sigma_n^2 >> \sigma_x^2$ so $\sigma_e^2 \approx \sigma_n^2$ and for high input SNR's $\sigma_n^2 << \sigma_x^2$ then we have $\sigma_e^2 \approx \sigma_{\varepsilon_x}^2$. As mentioned in Section 2.2, the AR process can model quasi stationary speech components far better than non-stationary components and onsets. So, during quasi stationary components (like vowels) the residual energy difference is very small for large filter orders. As we fixed the filter order to half of the maximum, the energy difference between original frame and extrapolated frame is different for different sounds. Figure 4 shows the scatter plot of the PG and the ED of noisy speech frames of the male and female speech utterances taken from TIMIT database at input SNR 7dB. The red spots represent all kinds of noisy frames, the blue ones are the frames which belong to the stop consonants of the female noisy speech signal, and green ones belong to the male noisy speech signal. As Figure 4 shows, almost all stop consonant frames are concentrated in an area where the prediction gain is less than 0 dB and the energy difference is more than 6 dB. As a conclusion we can classify the stationary and non-stationary frames by setting the thresholds of PG as $\tau_p = 0$ dB and of ED as $\tau_e = 6$ dB.

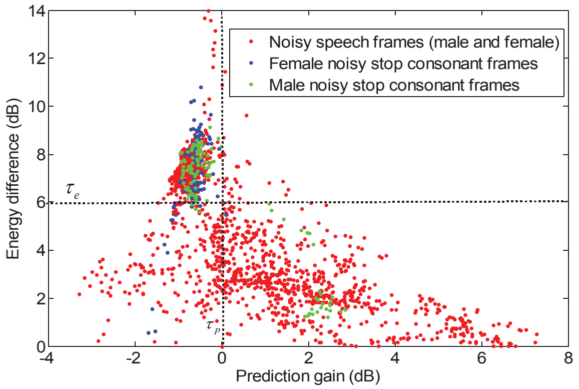

Fig. 4. Shows the relation between the prediction gain and the energy difference for different noisy speech frames.

The condition to classify the frame in the first part of the algorithm is as follows,

$$\begin{array}{rl} Non-stationary & \text{if } PG_i < \tau_p,\ ED_i > \tau_e \\ Stationary & \text{otherwise} \end{array}$$

The thresholds τ_p and τ_e are valid for all the input SNRs less than or equal to 7dB because as mentioned before as SNR decreases prediction gain also decreases and energy difference increases. For lower SNRs (< 7dB) the stop consonants shown in Fig. 4 further moves towards top-left corner. For the above 7dB input SNRs the thresholds changes in accordance with the prediction gain and the energy difference.

If the frame is classified as non-stationary frame it needs to be process further to be considered as stop consonant frame. Otherwise the algorithm ends at this step for that particular frame by setting it's frame index I_i to zero and starts again with the next frame.

3.2. *Step 2:* Periodicity check

As stated before signal modeling is very poor in case of non-stationary as well as during onsets of the recorded speech signal. Every onset in the speech signal may not be a stop consonant onset so in order to eliminate voiced onsets from non-stationary components further classification is required.

Voiced onsets consists of more or less constant frequency tones of some duration. Unvoiced speech is aperiodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract as when stop consonants are spoken. Voiced onsets can be identified and extracted by checking their periodic nature.

For checking the periodicity of the noisy frame, the time domain based autocorrelation pitch estimation method is implemented [7]. All voiced onset frames are eliminated from further processing in this step by setting their frame index I_i to zero. As stop consonants are unvoiced speech components, during periodicity check they would not exhibit any periodic behavior. After this step the only remaining classification is in between the noise and the stop consonant frames.

3.3. *Step 3:* Zero crossing rate (ZCR)

ZCR is one of the important and useful features for speech classification. The typical stop consonant waveform is shown in Fig. 5, where we can see the sudden burst followed by a slow decay. The useful way to detect stop consonants is to look for their transient region in the signal.

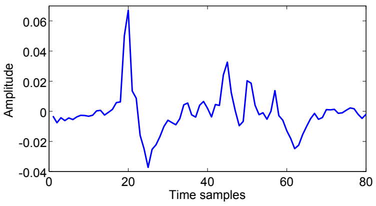

Fig. 5. Depicts the typical structure of the stop consonant waveform.

In this step, based on the structure of the stop consonants, the feature ZCR is used for final classification in between noise and speech stop consonant frames. The ZCR of the i^{th} frame Y_i is calculated according to equation (10).

$$ZCR_i = \frac{1}{2(N-1)} \sum_{n=2}^{N} |sgn(y_{iR+n}) - sgn(y_{iR+n-1})| \tag{10}$$

We analyzed the ZCR behavior of the stop consonant frames (SCF) along with the noise frames (NF). Found that the ZCR of the SCFs are different than NFs. In case of white noise NFs higher ZCR than SCFs because of their frequent signal sign change and also because of the slow decay region in Stop consonant. In case of babble noise the ZCR of

SCFs shows higher than compared to NFs. As explained, the thresholds of the final decision (ZCR based decision) depends on the type of input noise.

4. RESULTS AND DISCUSSION

The system was tested using 700 utterances for 60 speakers from different dialects of the TIMIT database mixed with additive white Gaussian noise, babble noise, and cafeteria noise at sampling frequency 16kHz. All noisy speech untterences are segmented with the frame length 512 and frame advance 64. Combining the extrapolation, periodicity and ZCR features to perform stop consonant detection for three different noise types yields the results shown in Fig. 6.

Fig. 6(a) shows true detection rate performance curves for three noise types and the Fig. 6(b) shows false alarm rate curves. True detection is counted when the noisy frame is detected as stop consonant frame and is also marked as stop consonant frame according to the TIMIT labeling. The true detection rate (TDR) is computed as the ratio of true detections to the number of stop consonants occurrence. The false alarm rate (FAR) is computed as the ratio of the false detections to the total number of false phoneme occurrences.

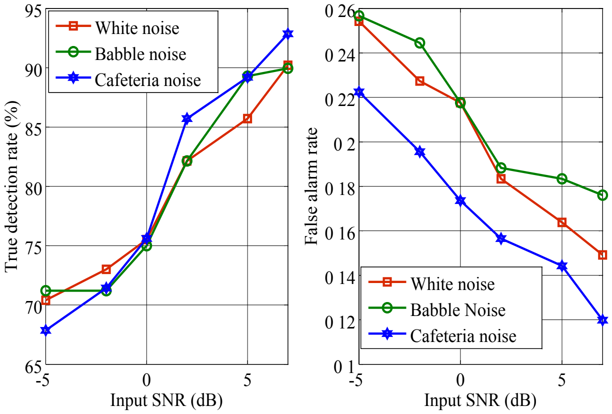

Fig. 6. The performance of the stop consonants detection in terms of TDR and FAR in case of three different noise types.

As shown in Fig. 6, at -5 dB input SNR the TDR and FAR achieved are 71% and 25.14% respectively and at 7 dB input SNR the TDR and FAR achieved are 90.21% and 10.81% respectively in case of white Gaussian noise. The thresholds of the speech features used in the detection algorithm can be modified according to the FAR and TDR requirements in applications.

5. CONCLUSIONS

In this paper we present an algorithm for stop consonant detection in continuous noisy speech at lower SNR's. The extrapolation technique was investigated along with AR process model order selection criteria. To achieve the aim, the chosen speech features are effectively used in the detection process of the stop consonants. The performance of the detector algorithm is demonstrated at lower SNRs in the range of -5dB to 7dB for different types of noises with a detection rate in the range of about 70% to 90% and with a false alarm rate in the range of 25% to 10%. By using adaptive thresholds instead of fixed thresholds for detection, we may reduce the false alarm rate further and increases the detection performance. The fixed thresholds could be adapted to the input SNR which needs to be estimated for this purpose. This will be investigated in future works.

6. REFERENCES

[1] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects," *Language and Speech*, vol. 43, pp. 273–294, 2000.

[2] D. Mauler and R. Martin, "Improved reproduction of stops in noise reduction systems with adaptive windows and nonstationarity detection," *EURASIP J. Adv. Sig. Proc.*, vol. 2009, 2009.

[3] A. Ali, J. Van Der Spiegel, and P. Mueller, "Robust classification of stop consonants using auditory-based speech processing," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 81–84, 2001.

[4] I. Kauppinen and K. Roth, "Improved noise reduction in audio signals using spectral resolution enhancement with time-domain signal extrapolation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1210–1216, 2005.

[5] C. W. J. Granger and M. J. Morris, "Time series modelling and interpretation," *Journal of the Royal Statistical Society*, vol. 139, no. 2, pp. 246–257, Feb 1976.

[6] P.M.T. Broersen, *Automatic autocorrelation and spectral analysis*, Springer-Verlag, 2006.

[7] H. Bořil and P. Pollák, "Direct time domain fundamental frequency estimation of speech in noisy conditions," in *Proc. EUSIPCO*, Vienna, Austria, 2004, vol. 2, pp. 1003 – 1006.