Voicing Features for Robust Speech Detection

Trausti Kristjansson, Sabine Deligne, Peder Olsen

IBM T.J. Watson Research
Yorktown Heights, NY, 10601
{tkristj,deligne,pederao}@us.ibm.com

Abstract

Accurate speech activity detection is a challenging problem in the car environment where high background noise and high amplitude transient sounds are common. We investigate a number of features that are designed for capturing the harmonic structure of speech. We evaluate separately three important characteristics of these features: 1) discriminative power 2) robustness to greatly varying SNR and channel characteristics and 3) performance when used in conjunction with MFCC features. We propose a new features, the Windowed Autocorrelation Lag Energy (WALE) which has desirable properties.

1. Introduction

Speech-silence discrimination and end-pointing important components of many speech recognitions systems.

Speech-silence discrimination is a challenging problem in the car environment where it is common to have high intensity semi-stationary background noise and high amplitude transient noises such as road bumps, wiper noise, door slams, tapping etc. High SNR conditions are also commonly encountered, such as when the car is stationary. We are therefore interested in features that are highly discriminative while being very robust to different conditions.

In this paper we focus on the inherent performance of the features that should be independent of the higher level decision mechanism. Various decision mechanism have been proposed such as likelihood ratio[1], HMMs [2] and hierarchical HMMs[3].

A simple and effective feature for speech detection in high SNR conditions is signal energy. Any robust decision mechanism based on energy must adapt to the relative signal and noise levels and the overall gain of the signal. In contrast, all the features reviewed in this paper are gain invariant.

MFCC features are also effective for discriminating speech from other environmental sounds although they were not designed for this purpose. In particular, the Mel filter removes the characteristics of excitation signal. For voiced sounds the excitation signal is periodic glottal pulse train signal, which manifests itself as harmonic structure in the spectrum (see Figure 1(b)).

Since MFCC features do not capture the harmonic structure of speech, an avenue of exploration is to extend the feature space with features that succinctly capture the strength of voicing of the signal.

An additional motivation for pursuing the structure of voiced speech rather than that of unvoiced speech is that in the car environment, unvoiced speech sounds are easily confusable with wind, road and fan noise.

2. Features for Voicing Detection

We investigated well known features and some recently introduced features. These features are:

- Autocorrelation Peak Count
- Spectral Entropy
- Maximum LPC Residual Autocorrelation Peak
- Spectral Autocorrelation Peak Valley Ratio
- Maximum Autocorrelation Peak
- Maximum Cepstral Coefficient

In addition, we introduce an extension of the Maximum Autocorrelation coefficient that is designed to improve its robustness, i.e. the:

- Windowed Autocorrelation Lag Energy

These features use either the autocorrelation or the spectrum or a combination in conjunction with an arbitrary nonlinear method for extracting a single measure. They all attempt to condense the harmonic structure of voiced speech into a single coefficient that is relatively efficient to compute.

2.1. Autocorrelation Based Features

A number of techniques in the literature are based on the autocorrelation of the signal [3, 4].

The periodic characteristic of speech signal makes it a good candidate for searching for self-similarity, i.e. repetitions of the filtered glottal pulse. However, the autocorrelation captures any repetitive signal, including motor noise.

The standard un-normalized autocorrelation is

$$a_j[k] = \sum_{n=k}^{N} x_j[n]x_j[n-k] \quad (1)$$

where x_j is the j-th segment of the signal and k is the lag. The autocorrelation can be normalized in a number of ways. In the lag-zero normalized autocorrelation each lag is divided by a[0]. This ensures gain invariance[4].

The short-time normalized autocorrelation normalizes each element by the energy of that lag. Hence it normalizes both for the number of lags, and the energy.

$$acorr_j[k] = \frac{\sum_{n=k}^{N} x_j[n]x_j[n-k]}{(\sum_{n=1}^{N-k} x_j[n]^2)^{\frac{1}{2}}(\sum_{n=k}^{N} x_j[n]^2)^{\frac{1}{2}})} \quad (2)$$

The methods based on the autocorrelation include the *Maximum Autocorrelation Peak*[4] which finds the magnitude or power of the maximum peak within the range of lags that correspond to the range of fundamental frequencies of male and

female voices. In our experiments, we used a range of 50Hz-400Hz corresponding to 320-40 lags respectively at a 16kHz sampling rate. Another measure is the *Autocorrelation Peak Count*[3] or the number of peaks found in a range of lags.

For pitch estimation in high SNR conditions, it is advantageous to remove the correlations of the vocal tract to reveal an approximation the glottal pulse train. This can be done by inverse LPC filtering. The *Maximum LPC Residual Autocorrelation Peak*[5] measure is based on finding the peak of the autocorrelation of LPC residual signal.

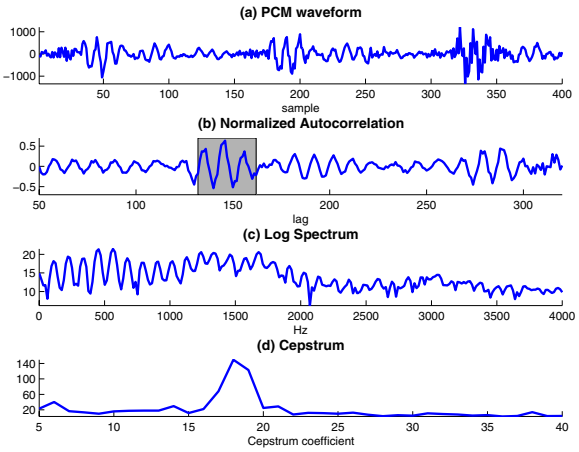

Figure 1: (a) The PCM waveform of a segment of voiced speech. Three glottal periods are shown. (b) The Normalized Autocorrelation shows a distinctive 'saw' pattern centered at the glottal period (lag 146 $\approx$ 109Hz fundamental at 16kHz sampling rate). The shaded box corresponds to a window of 30 lags. The energy of the autocorrelation lags in this window corresponds to the WALE coefficient for the speech segment. (c) The Log Spectrum shows regularly spaced harmonic peaks characteristic of voiced speech. (d) The Cepstrum has a very distinct peak corresponding to the fundamental frequency.

2.2. Spectrum Based features

The *Spectral Entropy* [6, 3] measure is found by interpreting the short-time spectrum as a probability distribution over a single discrete random variable X and then calculating the entropy of the distribution. The *spectral distribution* is found by normalizing the values of the short-time spectrum $p_X(f) = \frac{s(f)}{\sum_{k=1}^{N} s(k)}$ where $s(f)$ is the spectral energy for frequency f, and p_X is the spectral distribution. Now we can calculate the spectral entropy for frame j as

$$H(j) = -\sum_{k=1}^{N} p_{X_j}(k) \log(p_{X_j}(k)) \quad (3)$$

Due to the harmonic structure of voiced speech, it is expected that voiced speech will have relatively low entropy while stationary background noise is expected to have high entropy. This tendency can be seen in Figure 2. Various noise signals are also expected to have low entropy such as alarms, brake squeaks and sirens.

The *Spectral Autocorrelation Peak Valley Ratio* (SAPVR) [7] measure was introduce in the context of *usable speech detection*. If the a single speaker is speaking, the spectrum will have regularly spaced harmonic peaks. If two speakers are speaking simultaneously, this structure will be distorted. SAPVR takes the autocorrelation of the magnitude spectrum to detect the harmonic regularity. After this operation the maximal ratio between first valley and second peak in the autocorrelation is found[1].

The *Cepstral Peak* has been used for pitch estimation [9] as well as voice activity detection. The cepstrum is computed as

$$ceps = DCT(log(|FFT(x)|^2)) \quad (4)$$

where x is a short segment of the signal[2]. It is well known that the low order cepstra characterize the vocal tract filter, whereas the higher capture the excitation. Figure 1(c) shows a clear peak corresponding to the excitation period.

In order to better capture the peak, we ran a difference operator over the cepstra, and then found the difference between the maximum value and the minimum value. This produced better results than directly using the maximum.

3. Windowed Autocorrelation Lag Energy

The *Windowed Autocorrelation Lag Energy* (WALE) measure is designed as a robust extension of the Autocorrelation Maximum Peak Amplitude metric.

Voiced speech is produced when the vocal cords produce a glottal pulse train that is then filtered by the vocal tract. The autocorrelation of a pulse train has a single peak at the lag corresponding to the period of the pulse train, which motivates the use of the Maximum Autocorrelation as a voicing indicator.

However, the vocal tract introduces correlations and spreads out the energy somewhat. The signal decays rapidly after the glottal pulse and energy is concentrated in that region. In the autocorrelation, this manifests itself as a 'saw' pattern seen in Figure 1(b). The motivation for the Windowed Autocorrelation Lag Energy is to better capture this structure by taking into account a short window where most of the energy should be concentrated when the signal is voiced speech.

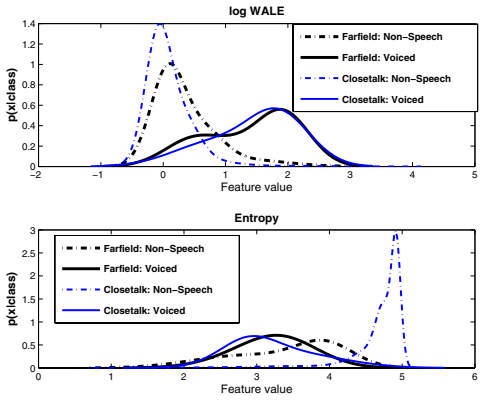

Figure 2: Distributions for voiced speech and background noise for the features log WALE and Spectral Entropy. Distributions are shown under the closetalk and farfield conditions. Notice that the distribution for background noise changes considerably between conditions for the entropy feature while log WALE is robust.

[1] a few variants are reported in the literature[7, 8].

[2] Note that unlike the Mel Filtered Cepstral Coefficient (MFCC) transform, it is important not to use a Mel Filtering or warping.

To calculate this feature, we slide the window across the autocorrelation lags, and calculate the energy of the lags in the window, at each shift point. The maximum value is then returned. A window of length 30 is shown as the shaded area in Figure 1(b), centered at the maximum value. Hence WALE is computed as

$$WALE(j) = \max_{l} \sum_{i=l}^{l+W-1} |acorr_j(i)|^2 \qquad (5)$$

where $acorr_j$ is the vector of autocorrelation coefficients calculated in Equation 2, W is the length of the lag window. In our experiments W was set to 15 lags. Notice that when the window is of length $W = 1$, WALE is equivalent to finding the square of Max Autocorrelation.

To further improve robustness, we can take advantage of the fact that voiced segments usually span a few consecutive frames. Since the voicing period will not change dramatically between consecutive frames, the maximum will be close in consecutive frames. We therefore define the Multi-Frame WALE ($WALE_{MF}$) as

$$WALE_{MF}(j) = \max_{l} \sum_{i=l}^{l+W-1} \sum_{t=j-\alpha}^{j+\beta} |acorr_t(i)|^2. \qquad (6)$$

where α and β designate how many past frames and future frames to consider, respectively. α and β were set to 1 in our experiments. When using Gaussian Mixture Models to model feature distributions, it is advantageous to use $\log(WALE)$ and $\log(WALE_{MF})$ instead.

4. Feature Evaluation

Three performance characteristics of the features are of particular interest to us: 1) their *discriminative power*, 2) their *robustness* to different conditions 3) how well they complement the MFCC features that are known to be effective features.

To assess these aspects of the features we collected a data-set consisting of 3 male speakers and 3 female speakers. Each speaker contributed about 10 minutes of speech data for a total of about 1 hour of data. The data was collected in different cars and a variety of conditions. An attempt was made to produce the whole range of noise conditions, transient sounds and situations where speech detection might fail. As an example, the data contains speech recorded when driving over road seams, washboards, potholes and other rough surfaces. Data was also collected by the roadside with door slams, trunk slams and with open windows and trucks driving by at high speeds in heavy rain. The data was collected on two channels where one channel recorded a far-field microphone mounted on the rear view mirror and the second channel was of a head mounted close-talking noise canceling microphone.

A portion of the close-talking data was hand labeled with three tags: *voiced, unvoiced* and *non-speech*. In order to get a similar labeling for the whole data-set, we ran forced alignment with a speech recognition system and known transcriptions on all the close-talking data. Phone models that correspond well to voiced speech segments were then used as a ground truth for voiced sounds. The recognition system also labeled non-speech segments reliably. The labeling was used as the ground truth for non-speech segments.

4.1. Discriminative Power

To assess the discriminative power of individual features we calculated two metrics of each feature when used alone; 1) the Symmetric KL distance between voiced and non-speech models in two noise and channel conditions, and 2) ROC curves showing Segment False Accept and False Reject error rates.

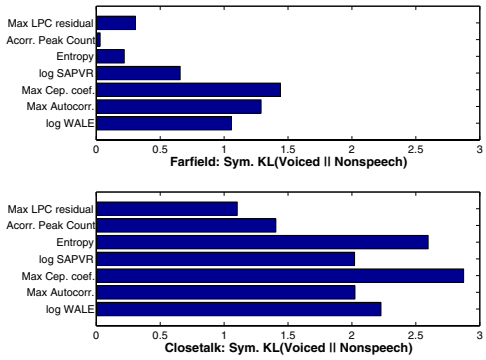

Figure 3: Symmetric KL distance between voiced distribution and non-speech distribution for far-field and close-talk conditions. High values are indicative of highly discriminative features.

Figure 3 shows the symmetric KL distance between $p(x|\text{voiced})$ and $p(x|\text{non-speech})$ for all features in the far-field condition and close-talk conditions. Note that in the close-talk condition, all features performed well. In the far-field condition, the performance of some of the features decreases considerably (e.g. Autocorrelation Peak Count and Spectral Entropy), when the distributions for speech and non-speech overlap. The cepstrum and the Max Autocorrelation and log WALE continue to perform well.

4.2. Robustness

To assess the robustness of the features we also calculated the Symmetric KL distance of equivalent distributions between the far-field and the close-talking condition. This measure gives an indication of how the features vary between two common conditions, i.e. the high SNR close-talking condition and a low SNR noisy far-field condition, and hence how they can be expected to perform in new unseen conditions.

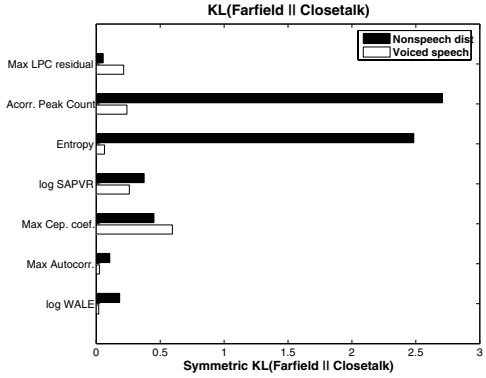

Figure 4: Symmetric KL distance between close-talk and far-field conditions. A large value is indicative of a non-robust feature.

False Reject Rate	5.00%	10.00%	15.00%	20.00%	24.62%	30.00%	40.00%	
FA MFCC only	68.28%	42.33%	25.19%	13.45%	7.33%	3.55%	1.41%	Average.
Max LPC residual	0.71%	4.52%	8.85%	15.63%	16.38%	8.00%	3.58%	8.24%
Acorr. Peak Count	0.06%	2.24%	1.34%	-1.61%	-0.81%	0.16%	-1.84%	-0.06%
Entropy	0.56%	5.04%	1.80%	-4.50%	-3.01%	-5.38%	-12.23%	-2.53%
log SAPVR	-0.48%	-0.39%	-3.15%	-3.31%	-2.04%	-7.02%	-14.21%	-4.37%
Max Cep. coef.	0.96%	-0.97%	-3.07%	-6.30%	-13.05%	-18.19%	-19.26%	-8.55%
Max Autocorr.	-2.71%	-0.52%	-8.08%	-11.88%	-21.18%	-21.49%	-17.94%	-11.97%
log WALE	-1.38%	-6.33%	-12.15%	-11.12%	-20.13%	-19.67%	-19.72%	-12.93%

Table 1: The table shows the relative percent change in False Accept rates when a single feature is added to the MFCC features. The numbers represent different points on an ROC curve. The baseline False Accept rate for the MFCC features is shown in the second row. Notice that the Max Autocorr feature and the log WALE feature perform well, where the WALE feature performs best when a low False Reject rate is desirable. The False Reject rate of 24.62% corresponds to a symmetric loss function of a Bayes classifier (i.e. cost of rejecting voiced speech is equal to cost of accepting noise).

Figure 4 shows that noise distributions for Autocorrelation Peak Count and Spectral Entropy change considerably, while Max Autocorrelation and log WALE show relatively good robustness characteristics for both the noise and voiced speech distributions.

4.3. Complement to MFCC features

To assess how well these feature complement the MFCC features which were used in our baseline Speech Detection system, we appended each of the features in turn to the 13 MFCC features and noted the affect on False Alarm rate at a particular False Reject rate. The voiced speech and noise models were trained on a large data-set consisting of in-car speech recorded at 0mph, 30mph and 60mph. Each feature was modeled with a 64 Gaussian mixture model. The models were combined assuming independence between the features. The test-set was the 1-hour far-field test set mentioned in the experiments above. The relative amount of low-SNR conditions and transient noises in the test set was larger than in the training set.

Table 1 shows the effect of adding each feature in turn. The numbers represent different points on an ROC curve. Different False Rejection rates were achieved by artificially biasing the prior probabilities of speech and noise models. In our application, the cost of missing a speech vector is high and it is desirable to select a low False Rejection rate.

The best improvement is achieved by selecting the features at the bottom of the table. The log-WALE feature is slightly better on average than the Max Autocorrelation feature. At low False Reject rates, the log-WALE feature outperforms the Max Autocorrelation feature.

It is interesting to note that inverse LPC filtering the signal prior to using the Max Autocorrelation is harmful to performance. This may be due to this feature not being robust to the types of noises in the test set that were not seen in the training set. It is also interesting to note that adding any of these features helps less when we bias towards low False Reject rates.

5. Discussion

We have evaluated a number of well-known voice activity features as well as some features that have recently been proposed. Our evaluation focused on performance in very difficult noise conditions and the robustness to different noise and channel conditions. We also introduced the Windowed Autocorrelation Lag Energy feature that has advantages over the Maximum Autocorrelation feature when low false reject rates are desirable.

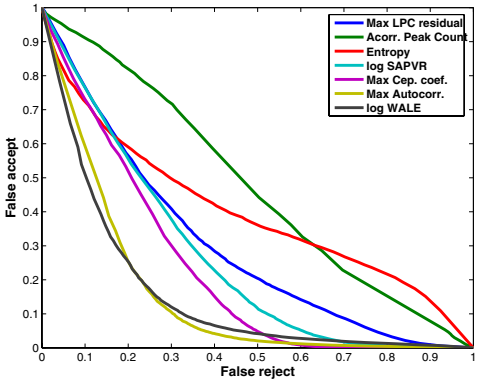

Figure 5: ROC curves for individual features.

6. References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 7, 1999.

[2] R. Sarikaya and J. H. Hansen, "Robust detection of speech activity in the presence of noise," in *Proc. Inter. Conf. on Spoken Language Processing*, 1998.

[3] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2003)*, 2003, pp. 816–819.

[4] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recogntion in noisy environments: The 2001 ibm spine evaluation system," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, 2001.

[5] W. Hess, *Pitch Determination of Speech Signals*. Springer Verlag, 1983.

[6] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. International Conference on Spoken Language Processing*, Sidney, Australia, Nov.-Dec. 1998.

[7] K. R. Krishnamachari and R. E. Yantorno, "Spectral autocorrelation ration as a usability measure of speech segments under co-channel conditions," in *IEEE Symposium on Intelligent Signal Processing and Communication systems*, 2000.

[8] R. E. Yantorno, K. R. Krishnamachari, and J. M. Lovekin, "The spectral autocorrelation peak valley ratio (sapvr) - a usable speech measure employed as a co-channel detection system," in *IEEE International Workshop on Intelligent Signal Processing*, 2001.

[9] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–337, 1999.