A ROBUST SPEECH DETECTION ALGORITHM FOR SPEECH ACTIVATED HANDS-FREE APPLICATIONS

D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador and X. Menendez-Pidal

Spoken Language Technology, Sony US Research Laboratories
3300 Zanker Road, San Jose, CA 95134, USA

ABSTRACT

This paper describes a novel noise robust speech detection algorithm that can operate reliably in severe car noisy conditions. High performance has been obtained with the following techniques: (1) noise suppression based on principal component analysis for pre-processing, (2) robust endpoint detection using dynamic parameters [1] and (3) speech verification using periodicity of voiced signals with harmonic enhancement. Noise suppression improves the SNR as compared with nonlinear spectrum subtraction by about 20 dB. This makes the endpoint detection operate reliably in SNRs down to –10 dB. In car environments, road bump noises are problematic for speech detectors causing mis-detection errors. Speech verification helps to remove these errors. This technology is being used in Sony car navigation products.

1. INTRODUCTION

Hands-free operation is a very important feature for speech activated systems. Speech detection provides a way to solve the problem for isolated word speech recognition. In addition, it has been shown that endpoint detection improves the isolated word recognition accuracy [2].

Many speech detection algorithms have been proposed [2]. For applications in car environments, a good speech detector should be noise robust, accurate and capable of real-time implementation. Good performance has been reported for moderate noise conditions such as SNRs above 5 dB. However, for severe noise conditions such as SNRs down to –10 dB in some car environments, no successful detection algorithm has yet been reported. In car environments, road bump noises are a frequent noise source. Most proposed speech detection methods use signal energy as the primary detection parameter. For these detectors, strong background noise such as road bump noise may be incorrectly detected as speech.

To obtain reliable speech detection for car applications, we proposed a speech detection algorithm that consists of three major parts: noise suppression, robust endpoint detection and speech verification. The noise suppression module is used as pre-processing for robust endpoint detection to suppress background noise. The noise suppressed signal is then passed to the robust endpoint detection module in which boundaries of utterances are detected. Since strong non-speech signals may also be detected as speech, the speech verification module performs verification for the signal between the boundaries and outputs the endpoints to a recognizer.

2. NOISE SUPPRESSION

2.1 Problem

Noise suppression is an important step for speech detection to operate in severe noise conditions such as SNRs down to –10 dB. Figure 1 shows a speech segment waveform (2a), noisy waveform (2b) corrupted by car noise at SNR –10 dB, and its short-time energy (2c). In Figures 2b and 2c, no obvious evidence of speech can be observed. Detectors designed for moderate noise conditions will perform poorly in these severe noisy conditions.

The performance of the speech endpoint detection algorithm depends on both the SNRs and the smoothness of parameters used. Low SNRs causes the detection failure rate to increase and a rough parameter curve corrupted by noise makes accurate endpoint detection difficult. To improve the performance of the speech endpoint detection algorithm in a low SNR, the SNR should be increased, and while the noise element with a large variance should be suppressed. The proposed method attempts to solve the above problem for speech detection under severe noisy conditions.

2.2 Subspace Method

In the proposed endpoint detection algorithm described in section 3, the primary parameter used to detect endpoints is the summation of each band output energy (or delta energy) of a filter-bank. The bands with large energy output dominate the overall SNR value. For a low SNR, these bands may not have a high SNR, since noise energy could be high in these bands. To have a high overall SNR, the energy from the bands that have a high SNR should be more heavily weighted. In other words, the weights should be directly proportional to the SNR of bands.

The Karhunen-Loeve transformation can be used to enhance this procedure, since feature data are projected onto the subspace on which the variances of noise data are maximized or minimized in its principal directions.

Let $\mathbf{n}$ denote the non-correlated additive random noise vector, $\mathbf{s}$ be the random speech feature vector and $\mathbf{y}$ stand for the random noisy speech feature vector, all with dimension p. Then $\mathbf{y} = \mathbf{s} + \mathbf{n}$. Assume $E[\mathbf{n}] = 0$, where E is the statistical expectation operator. If $\mathbf{n}$ has a nonzero mean, the mean is simply subtracted from $\mathbf{n}$ before analysis. The correlation matrix of noise vector can be expressed as $\mathbf{R} = E[\mathbf{n}\mathbf{n}^T]$.

R has its singular value decomposition expressed as

$$\mathbf{R} = \mathbf{V}\,[\mathrm{diag}\,\lambda]\,\mathbf{V}^T \qquad (1)$$

where **V** is a p-by-p orthogonal matrix in the sense that its column vectors (i.e., the eigenvectors of **R**) satisfy the conditions of orthonormality and λ is a p-by-1 vector defined by the eigenvalues of **R**, $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, ..., \lambda_{p-1}]^T$.

Since each eigenvalue of **R** is equal to the variance of projection data in its corresponding principal direction, with zero means, vector $\boldsymbol{\lambda}$ also defines the average power vector of projection data.

Let **q** denote the average power vector of the random speech projection vector by **V**

$$\mathbf{q} = [\beta_0, \beta_1, ..., \beta_{p-1}]^T \qquad (2)$$

Then SNR r_i for element (or band) i is given as

$$r_i = \beta_\iota / \lambda_\iota \text{ , i=0, 1, ..., p-1.} \qquad (3)$$

A simple way to have a weight vector **w** whose element values are directly proportional to the SNR is to have

$$w_i = (r_i)^\alpha \text{ , i=0,1,…p-1,} \qquad (4)$$

where α is a constant. Since vector **q** is not available in noisy environments, to calculate vector **w**, we may use vector **q'** which is estimated from the noisy speech vector **y** and noise vector **n**. For simplicity, currently we set **q** to the unit vector and α to 1. With this setting, the weight vector can be expressed as

$$w_i = (1/\lambda_\iota) \text{ , i=0,1,…p-1,} \qquad (5)$$

which can be explained that high noise bands are lightly weighted and low noise bands are heavily weighed.

2.3 Implementation

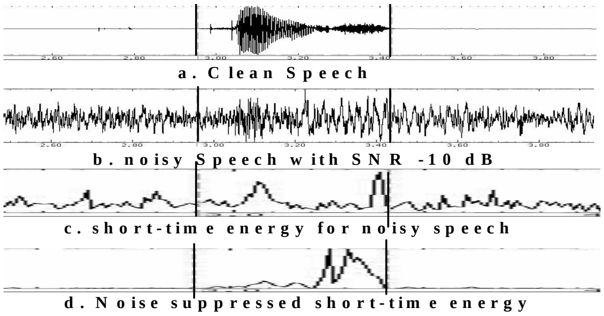
The implementation of the subspace method has three steps. 1) Calculate eigenvectors **V** and eigenvalue vector $\boldsymbol{\lambda}$ of the correlation matrix of background noise data and set w_I using equation (5). 2) Project the noisy speech vector **y** onto the subspace spanned by **V** with $\mathbf{y_s} = \mathbf{V^T y}$, where $\mathbf{y_s}$ represents the projection vector of **y**. 3) Weigh the projection vector $\mathbf{y_s}$ by **w** to form output vector **z** with $z_i = y'_i w_i$, i=0,1,…p-1. Figure 2d shows the noise suppressed short-time energy of noisy speech shown in Figure 2b. Clear boundaries of the utterance are observed.

3. ROBUST ENDPOINT DETECTION [1]

The endpoint detection uses dynamic features and reliable adaptive thresholds contingent upon local Signal-to-Noise Ratios (SNR). The algorithm employs a two-step search scheme [2]: reliable island search and boundary refinement.

3.1 Parameters

Delta short-term energy (hereafter called the dynamic time-frequency (DTF) parameter) is used as parameters to detect the endpoints. The DTF parameters are calculated with the equation given below,



a. Clean Speech

b. noisy Speech with SNR -10 dB

c. short-time energy for noisy speech

d. Noise suppressed short-time energy

Figure 1. An example to show noise suppression performance

$$DTF'(i) = \sum_m \left| \sum_{l=1}^{2} l(y_{i+l}(m) - y_{i-l}(m)) \right| / 10 \qquad (6)$$

where $y_i(m)$ is the m-th channel energy of output feature vector from the noise suppression section at time instant i (in frame). DTF"(i) is then smoothed by a 5-point median filter to obtain DTF(i). It is observed that the DTF is more robust against noise than the signal energy for car noise.

3.2 Adaptive Threshold Determination

There are four thresholds, one to detect the beginning point of reliable islands (T_{sr}), another to detect the ending point of reliable islands (T_{er}), one more to refine the beginning point (T_s) and the final one to refine the end point (T_e). Thresholds should be adaptive both to the background noise and to the signal-noise-ratio (SNR).

For real-time implementation, only the local SNR is available. Let SNR_{le} and SNR_{le} be the SNR for the beginning and ending points respectively. With background noise N_{bg}, T_s and T_e can be determined with the equations given below,

$$T_s = N_{bg}\sqrt{(1 + SNR_{ls}^{\;2} / c_s)} \qquad (7a)$$

$$T_e = N_{bg}\sqrt{(1 + SNR_{le}^{\;2} / c_e)} \qquad (7b)$$

where c_s is a constant for the beginning point determination and c_e is a constant for the ending point determination. T_{sr} and T_{er} can be determined in a similar way. N_{bg}, SNRs and the thresholds are updated as the search progresses.

3.3 The Algorithm

The algorithm consists of two steps: reliable island detection and boundary refinement. The beginning point of the reliable island is detected when DTF(i) is first over T_{sr} for at least 5 frames and the ending point of the reliable island is detected when DTF(i) is below T_{er} for at least 60 frames (or 600 ms) or T_e for at least 40 frames (400 ms). After the beginning point of the reliable island is detected, a backward-searching (or refinement) procedure is used to find the beginning point of the utterance. The searching range is limited to 35 frames (or 350 ms) from the beginning point of the reliable island. The beginning point is found when DTF(i) is below T_s for at least 7 frames. A similar procedure is applied to find the ending point of the utterance.

4. SPEECH VERIFICATION

The speech verification method uses the harmonics of the fundamental frequency F_0 of voiced signals to determine whether the input signal corresponds to an utterance. There are three major steps in this method. 1) The first step carries out the harmonic enhancement by summing adjacent frames of short-time spectra. 2) Pitch detection is then implemented by spectrum comb analysis. 3)The final step calculates the confidence measure of voice quality based on the magnitude and the peak sharpness.

4.1 Pre-processing

Pre-processing performs down-sampling from 16 kHz sampling rate to 4 kHz sampling rate with a 0-2000 Hz bandwidth. A 1024 point FFT is applied to each 40 ms Hanning windowed data, shifted at the rate of one frame every 10 ms.

4.2 Harmonic Enhancement

Noise suppression is implemented by summing N adjacent frames of spectra. Specifically, assume that corrupted noise is additive and let *$Y_i(k)$* denote the noisy spectrum at frame i. Then

$$Y_i(k) = S_i(k) + N_i(k),$$

where *$S_i(k)$* is the speech spectrum at frame *i* and *$N_i(k)$* is the noise spectrum at frame *i*. Summation *$Z_i(k)$* can be expressed as

$$Z_i(k) = \sum_{l=0}^{N-1} S_{i-l}(k) + \sum_{l=0}^{N-1} N_{i-l}(k), \tag{8}$$

where *N* is the number of frames for summation. Considering only the fundamental frequency and its harmonics, and knowing that for voiced signals (particularly vowels), within a short period, spectra for adjacent frames are similar, the summation of speech spectra *$Zs_i(k)$* can be approximated as

$$Zs_i(k) = \sum_{l=0}^{N-1} S_{i-l}(\beta_l k) = NS_i(k) \tag{9}$$

Here β_l represents the frequency scale to align the slightly different fundamental frequencies between the *(i-l)*-th frame spectrum and the *i*-th frame spectrum. Assuming that noise at each frame is not correlated with the speech and the noise of its adjacent frames, the SNR gain from the summation is

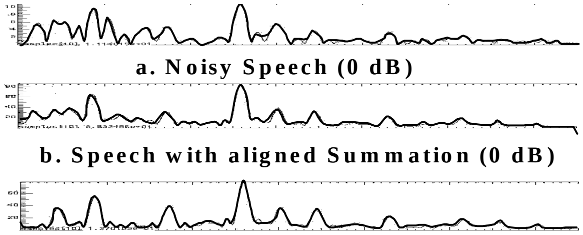
$$SNRg = 10\log_{10}(N). \tag{10}$$

β_l can be obtained with

$$\beta_l = \arg\min_a\left(\sum_k \left|S_i(k) - S_{i-l}(ak)\right|\right) \tag{11}$$

An exhaustive search can be used to find β_l within a small range, say [0.95 1.05] with a delta equal to 0.01.

Figure 2 shows an example with N=6. It can clearly be seen that noise has been successfully suppressed as shown in Figure 3b.



a. Noisy Speech (0 dB)

b. Speech with aligned Summation (0 dB)

c. Clean Speech

Figure 2. Comparison for Noise Suppression used for speech verification

4.3 Pitch Detection

Spectrum comb analysis [3] is used to detect the pitch from the noise-suppressed spectrum. Spectrum comb analysis is a computationally efficient method, classified as one of the second generation methods [4] which are shown to be especially resistant to noise. The method can be expressed as a correlation between a teeth window and a spectrum. For voiced signals, the frequency at which the maximum peak locates is considered as the fundamental frequency.

4.4 Confidence Measure

The confidence measure or voiced/voiceless classification is based on the results from the Pitch Detection section. One way [5] to make voiced/voiceless classification is to use the auto-correlation value at the maximum peak frequency. An alternative way is to directly use the results from the pitch detection by measuring the frequency quality of the maximum peak. It is clear that the sharper the peak, the better the signal in terms of closeness to the pure sine waveform. Two parameters are used: 1) Magnitude ratio defined as R= $(M_{peak} - M_{avg})/M_{peak}$, where M_{peak} denotes the maximum peak magnitude and M_{avg} is the average and; 2) Quality factor Q defined as the width between the half magnitude points from the maximum peak. The utterance is classified as speech if RQ > 0.05 consecutively for at least 4 frames.

5. EXPERIMENTS

5.1 Tasks and Criteria

Three tasks were investigated in the experiments. The first task is to evaluate the SNR improvement from the proposed noise suppression method over the conventional spectral subtraction method with full-wave rectification. An English database containing 13000 tokens, and noise data collected from a car running on streets and highways are used in this task. The second task is to evaluate the performance improvement of speech endpoint detection algorithm by using the proposed noise suppression method. An English database containing 10-speaker data, 5 females and 5 males, is used. The same noise described above is also used in this task. The final task is to evaluate the proposed speech verification method. Data recorded in a car driven on streets and highways is used for this task. The data has 118 isolated words contained in 1 hour and 17 minutes of noisy car data. The SNR ranges from 0 dB to 30 dB.

For the SNR evaluation, the SNR is defined as

$$SNR = 10\log_{10} \frac{\sum_{m=1}^{M}\left(\left|s(m)^{T}\right|w\right)}{\sum_{m=1}^{M}\left(\left|y(m)-s(m)\right|^{T} w\right)},$$

where **s**(m) denotes the projection speech vector, **y**(m) is the projection noisy speech vector, both at frame m, **w** is the weight vector.

For the endpoint detection evaluation, experiments were conducted to evaluate the performance of the speech detection algorithm with the criterion of average difference between the detected endpoints and hand-marked endpoints.

For speech verification, the endpoint detection was first carried out to find word boundaries on the noise-data. Later, for each utterance found with the endpoint detector, speech verification was carried out.

5.2 Results and Analysis

Results are given in Table I for task one and in Table II for task two. In task one, the SNR of original noisy speech was set to -10 dB. Noisy speech was pre-emphasized for high frequency components with α = 0.97 for the pre-emphasis case. The noisy speech was analyzed with a 24-band filter-bank. The energy of each band output forms the feature vectors.

Table I. SNR improvement for noisy speech of -10 dB SNR

	w/o pre-emphasis		with pre-emphasis	
	SNR	SNR Improve.	SNR	SNR
Consona	-10.17	18.27	-0.61	9.59
Vowel	-5.07	13.73	1.24	7.98
word	-6.04	18.91	3.77	10.0

Table II. Average Difference Comparison between the detector with (and without) noise suppression

		-10dB (ms)	-15dB (ms)	-20dB (ms)	-30dB (ms)
B	w/ NS	167	157	200	384
B	w/oNS	156	206	289	853
E	w/ NS	172	222	274	476
E	w/oNS	250	289	346	704
M w/NS		3	13	62	253
Mw/oNS		45	241	500	752

B – Beginning point; E – Ending point
M -- No. of missing tokens; w/ NS – with noise suppression;
w/o NS – without noise suppression

Three broad clusters, consonant, vowel and word are under evaluation. From the results, it can be seen that our proposed method improves the SNR by 10 dB for words, 7.89 dB for vowels and 9.59 for consonants over the spectral subtraction method with pre-emphasis, and 18.91 dB for words, 13.73 dB for vowels and 18.27 dB for consonants without pre-emphasis.

For task two, there are four SNR levels including -10 dB, -15 dB, -20 dB and -30 dB. A total of 860 tokens were used in this investigation. From these results, it is obvious that speech detection with noise suppression performs much better in terms of number of missing tokens and accuracy.

Results from task 3 are given in Table III and Table VI. The detector correctly found all 118 speech tokens, but incorrectly found 16 non-speech tokens as shown in Table 1.

Table III. Performance of speech verification (SP) algorithm in correction no. of tokens

	Speech Tokens	Mechanical Noise tokens	Road bump noise tokens	Human noise tokens
W/o SP	118	5	9	2
W/ SP	116	1	2	1

Table III shows that the number of noise tokens (mechanical, road bump and human) mis-detected as speech tokens by the endpoint detector is 16. This number is reduced to 4 by using speech verification, however, 2 actual speech tokens are still mistakenly re-classified as noise. Subsequently, the error rate is reduced from 11.94% to 4.48%.

6. CONCLUSION

A noise robust speech detection algorithm has been proposed which uses three technologies: noise suppression, robust endpoint detection and speech verification. The method was developed for the applications which have very low SNRs. Experiments shows that SNRs are greatly improved with the proposed noise suppression, which makes the endpoint detection operate reliably in SNRs down to –10 dB for car noises. For speech verification, experiments show that the method is robust to car noise. From the results, it can be concluded that the proposed method has good results for speech-activated hands-free systems such as cellular telephones or car navigation systems, particularly in the very low SNR conditions.

7. REFERENCES

[1] Wu D., M. Tanaka, R. Chen and L. Olorenshaw, "A Robust Endpoint Detection Algorithm for Speech Recognition in Cars" Proceedings-97 of Sony Research Forum, Tokyo, 1997.

[2] Junqua J. C. B. Mak, and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 3, Jul. 1994

[3] Martin P., "Comparison of Pitch Detection by Cepstrum and Spectrum Comb Analysis" ICASSP-82, pp. 180-183, 1982

[4] Dik J. Hermes, "Pitch Analysis" in Visual Representations of Speech Signals, Martin Cooke, Steve Beet and Malcolm Crawford (eds.), John Wiley, 1993

[5] Dik J. Hermes, "Measurement of Pitch by Subharmonics Summation" J. Acoust. Soc. Am. 83(1), January 1988.