ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Speech and Word Detection Algorithms for Hands-Free Applications

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

Spoken Language Technology, Sony US Research Laboratories
E-mail:duanpei@cisco.com, xavier@slt.sel.sony.com

ABSTRACT

This paper describes a robust speech detection algorithm for speech-activated hands-free applications. The system consists of three techniques: (1) noise suppression with efficient implementation, (2) robust endpoint detection and (3) speech verification using garbage modeling and confidence measure. With efficient implementation, noise suppression improves the SNR by roughly 10-20 dB. The endpoint detection uses the technique described in [1] with improvement for non-stationary noise. Garbage modeling and confidence measure are used to handle out-of-vocabulary (OOV) words and background pulse noise.

1 INTRODUCTION

Hands-free operation is a very important feature for speech activated systems. Speech detection under background noise provides a way to solve the problem for isolated word speech recognition.

In ICASSP-99, we proposed a speech detection algorithm for hands-free operation. The algorithm 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 stationary 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 a speech utterance with the previous two steps, the speech verification module performs verification for the signal between the boundaries.

In this algorithm, noise suppression based on the Karhunen-Loeve Transformation (KLT) greatly improves the SNR. Due to this technique, the system could be operated reliably in SNRs down to -10 dB. However, the algorithm requires a large amount of computation to perform KLT, which makes its application difficult for real-time implementation. The other key technique, robust speech verification based on harmonics of the voiced signal, works well for environment pulse noise without periodicity. However, for noise with periodicity such as speech signals, the algorithm will not work.

To solve these problems and to obtain reliable speech detection for hands-free applications, we propose an improved speech detection algorithm that again consists of three major parts: noise suppression, robust endpoint detection and speech verification, each of them performing the same function as before. The noise suppression is improved with efficient implementation. The major improvement of endpoint detection is to handle non-stationary environment noise. The previous verification method is replaced by garbage modeling for unknown speech/noise signals and confidence measures to verify that the signal between the boundaries is a speech token in the vocabulary.

2 NOISE SUPPRESSION

2.1 Problems

Noise suppression based on KLT proposed in our ICASSP-99 paper [2] obtained about 20dB SNR improvement for car noise. SNR improvement ranging from 10-20 dB has also been obtained for other types of noise from environments including OFFICE, CONFERENCE EXHIBITION HALL and STREET. However, the computation load required with KLT is an obstacle for the method to be used in real-time applications.

2.2 New noise suppression methods

As stated in our ICASSP-99 paper, we used short-term energy as the primary endpoint detection parameter that is the summation of output energy from each band of a filter-bank. Therefore, the bands with large energy output dominate the overall SNR value. However, 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 was used to enhance this procedure, since feature data were projected onto the subspace on which the variances of noise data were maximized or minimized in its principal directions.

There are three steps in implementing the noise suppression in the previous method [2]: (1) calculation of channel background noise energy vectors, (2) KLT on the background noise energy vectors and (3) weight vector calculation based on the SNRs in the projection subspace. To reduce the computational load, we simply omit the step of KLT and calculate the weight vector directly based on the energy vectors of channel background noise. Experiments show that the SNR is only slightly decreased after skipping step 2.

Let **n** denote the non-correlated additive random noise vector, **s** be the random speech feature vector and **y** stand for the random noisy speech feature vector, all with dimension p. Then **y** = **s** +

n. Let **q** denote the estimated average energy vector of the random speech vector **s**,

$$\mathbf{q} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^T, \tag{1a}$$

and **l** be the estimated average energy vector of background noise **n**,

$$\mathbf{l} = [\lambda_0, \lambda_1, \ldots, \lambda_{p-1}]^T. \tag{1b}$$

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

$$r_i = \beta_i \;/\; \lambda_i \qquad i=0, 1, \ldots, p-1, \tag{2}$$

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

$$w_i = (r_i)^{\alpha}, \; i=0,1,..p-1, \tag{3}$$

where α is a constant. Since vector **q** is not available in noisy environments, to calculate vector **w**, we may use vector $\mathbf{q}'$ 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

$$w_i = (1/\lambda_i), \; i=0,1,..p-1, \tag{4}$$

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

3 ROBUST ENDPOINT DETECTION

A similar procedure as described in [1] was used for the endpoint detection. The main modifications are described below.

3.1 Parameters

It was observed that better results were obtained with short-term energy for the new noise suppression method than Dynamic Time-Frequency (DTF) used before. Therefore, short-term energy is used as the detection parameter. The parameter for the i-th frame is calculated with the equation

$$e(i) = \sum_{m=0}^{M-1} y_i(m), \tag{5}$$

where $y_i(m)$ is the m-th channel energy at frame i and M is the number of channels of the filter-bank.

3.2 Utterance validity

In our previous method, when a signal has DTF parameters over the reliable island threshold Trs for P consecutive frames, the signal segment is detected as a valid utterance. For non-stationary noise, the system frequently detects noise as speech due to many (noises) pulses that last longer than P frames. To improve detection reliability, several constraints have been imposed onto the algorithm.

1) Multi-pulse for Reliable Island Detection*:* For multi-syllable words, the single syllable alone may not last long enough for the condition of P consecutive frames. Considering the problem, P is counted as the number of samples that are over Trs, within a limited time period.

2) Minimum Signal Power Constraint*:* When the signal is too soft (even if it satisfies condition 1), it should be classified as noise. A limited energy threshold is applied to eliminate signal segments that are too soft.

3) Duration Constraint*:* When the detected segment duration is rather small or rather large, the segment is unlikely to be an utterance. So, a maximum and a minimum duration are used to constrain the utterance to be recognized.

4) Minimum Signal Power Constraint for Short Utterances*:* Some word utterance may have a short duration. To distinguish it from background pulse noise, the short utterance should have high energy. So, another energy constraint is imposed on the short utterances.

4 SPEECH VERIFICATION

Noise environments generally degrade the performance of speech recognition systems. For hands-free isolated-word recognition systems, background speech and environment noise pulses are the main sources for performance degradation. Speech verification is designed to distinguish the words in the vocabulary from the words/noises out of vocabulary (OOV).

The proposed speech verification consists of two parts, garbage modeling and confidence measure. Garbage modeling is intended to model the known types of noise, while confidence measure handles unknown types of noise and the noise that is not feasible to model such as noises from TV programs.

4.1 Garbage modeling

Like modeling the words in vocabulary, the words/noises out-of-vocabulary could also be modeled by specific HMMs. Garbage modeling is done to incorporate these specific HMMs into a recognizer to identify the out-of-vocabulary speech words and the noise pulses frequently present in application environments.

4.2 Modeling the out-of-vocabulary words

For OOV speech words, a specific HMM was trained with the entire training database to cover the whole acoustic space. That is, all words in the training database are considered as garbage. The garbage model can be trained with the normal phone models or trained separately. It was observed that training the normal phone models together with the garbage model increases the recognition accuracy. During the training, the garbage model should be trained first, and then the phone models are trained.

The recognition system is built by combining the normal models and the garbage model. The dictionary is augmented with the garbage model word. In this way, the recognition engine can be used to recognize the words in vocabulary or to reject unknown (out-of-vocabulary) words if the garbage model is recognized.

4.3 Modeling the environment noises

Although it is almost impossible to count all kinds of noises, there are many types of noises that are frequently presented in application environments. Similar to modeling the out-of-vocabulary words, these noises can also be modeled with specific HMMs. For the Sony AIBO applications, 21 types of noises shown in Table 1 were considered.

Surface contact	put a plate/cup, walk, drag chair, close door, knock on doors
Human noise	cough, laugh, uh, um, mm, hum
AIBO contact noise	hit-head, push-switch, hit-body, lift-up-body, put-down-body
Others	keyboard click, paper wrinkle, key jingle, clap

Table 1. Noise list

Noises are first clustered into several clusters, each having similar acoustic characteristics. Each cluster is then trained with a single model to characterize those types of noises within the cluster. To conduct clustering, a clustering algorithm was developed. The algorithm uses the recognition scores (or likelihood probability) of one word to every other word as their distances. The algorithm starts with 21 clusters, one for each type of noise, and merges them to the required number of clusters according to their mutual distances [5].

As for the out-of-vocabulary words, models of environmental noises are combined with normal word models to build the recognition system.

4.4 Confidence Measure

Confidence measure uses the differential score (or distance) between the first and the second candidates. If the distance is larger than the pre-determined threshold, the token is rejected as garbage. Otherwise, it is accepted as a recognized token. To have good recognition accuracy, it is desirable to have a small threshold. However, a small threshold will yield a high failure rejection rate. Therefore, a trade-off is usually made between the recognition accuracy and the failure rejection rate.

Several ways to use the differential scores have been investigated. Performance depends on the number of thresholds used. One simple way, denoted as method 1, is to use a single threshold for all words in vocabulary. Since each word has a different statistical duration, long words have lower scores than short words. Therefore, normalization to duration was conducted before comparison.

For a given dictionary, since each word has a different statistical distance to every other word, a single threshold may not yield good results. One way, denoted as method 2, to increase the number of thresholds is to associate each word in the vocabulary with a threshold.

It is observed that among the vocabulary words, one word is usually confused with a few of words. Among these confused words, the distances are small. Therefore, to keep the recognition accuracy high, small thresholds must be chosen in method 2, which adversely increases the rejection failure rate. Method 2 is improved, denoted as method 3, by handling the confused words separately. Specifically, each word confused with the considered word is associated with a threshold and is not counted for the threshold determination for other words. Without these confused words, the threshold for non-confused words can be increased.

For a given dictionary, thresholds can be determined with a set of test utterances. For every word pair in the dictionary, we can have its minimum and its maximum distances for the given test utterances. These minimum and maximum distances, or values between the extremes, can be chosen as the thresholds. For the method with a single threshold, the minimum and maximum are chosen over all words. For the given test set, if the minimum distance is chosen as the associating threshold for the corresponding word in the dictionary, the rejection procedure will not degrade the system recognition accuracy. On the other hand, if the maximum distance is chosen as the associating threshold for the corresponding word that is wrongly recognized to the input token, the system will obtain the maximum rejection accuracy (100%). A threshold between these two extreme values will provide the trade-off performance between rejection and recognition accuracy. A single variable α ranging from 0 to 1 can be used to control the trade-off performance.

5 V. EXPERIMENTS AND RESULTS

5.1 Tasks and criteria

A series of experiments have been conducted for the proposed method. For noise suppression, the experiment is conducted to compare the SNR improvement between the new method and the subspace method described in [2] over the SNR using 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 is used in this experiment.

For speech verification, experiments were carried out to evaluate the performance of garbage modeling and confidence measure in terms of recognition accuracy and failure rates. All experiments used the same clean English database: the model training set, 167 speakers, ~26 hours data; test set A, 2260 tokens, 125 words (relatively long), 26 speakers; test set B, 586 tokens, 100 words (single words for AIBO), 30 speakers. Environmental noises include 21 types of noises as described above in section 4.3. There are a total of 1671 noise tokens, 1059 tokens for training and 612 tokens for test. Three measure matrices are used in evaluation.

1) Recognition Accuracy: RA=CV/TV, where CV is the number of correctly recognized tokens of words in the vocabulary, and TV is the total number of tokens of words in the vocabulary.

2) Garbage Noise Recognition Accuracy: GA = CG/TG, where CG is the number of correctly recognized noise garbage tokens and TG is the total number of noise garbage tokens.

3) Failure Rate: FR = FT/TT, where FT is the total number of tokens failed to reject and accept, and TT is the total number of tokens to be rejected including noise tokens, mis-recognized tokens of words in vocabulary, tokens of out-of-vocabulary words and tokens that are correctly recognized but failed to be accepted.

5.2 Results and analysis

For noise suppression, results are shown in Table 2. The SNR of original noisy speech was set to -10 dB. The noisy speech was analyzed with a 24-band filter-bank. The energy of each band output forms the feature vectors.

	NSS	SNS	DWNS
	SNR	**SNR Impro.**	**SNR Impro.**
Consonant	-12.11	19.19	17.61
Vowel	-6.71	12.61	12.61
Word	-	19.96	18.49

NSS -- nonlinear spectral subtraction
SNS -- subspace Noise Suppression
DWNS --Direct Weighting Noise Suppression (the new method)

Table 2. SNR improvement for noisy speech at -10 dB SNR

Three broad clusters, (consonant, vowel and word) are under evaluation. From the results, it can be seen that DWNS yields high SNR improvement over the NSS, although the improvement is not as good as that obtained using SNS.

The results for out-of-vocabulary word modeling are presented in Table 3. It can be seen that garbage modeling indeed provides rejection capability for the out-of-vocabulary words. However, the word recognition accuracy decreased from 99.0% to 93.3% with data Set A. For this reason, garbage modeling for out-of-vocabulary words was not combined with confidence measure.

	Word accuracy w/o rejection	Word Accuracy	Rejection Accuracy
Set A	99.0%	93.3%	58.8%
Set B	90.2%	87.7%	90.1%

Table 3. Evaluation of the phonetic HMM word accuracy and OOV-HMM rejection accuracy in the test sets A and B

Several experiments to evaluate the garbage and confidence measure have been conducted. The system was built by combining the normal models for words in vocabulary with noise garbage models (without out-of-vocabulary word models). The total number of Gaussians of normal models is 1127, which yields 97.57% of recognition accuracy for set A data with/without garbage models. Results are shown in Table 4 with all thresholds set to the minimum distances, which means that there are no failure acceptation errors for the given data sets.

From the results, it is clear that the confidence measure based on differential scores is effective. The performance is greatly improved by handling confused words separately.

Configuration	Acc.	Garbage Acc.	Failure Rate
one threshold all words	97.57	94.53	56.62
one threshold one word	97.57	94.53	18.36
One confused word	97.57	94.53	7.86
Two confused word	97.57	94.53	4.44

Table 4. Performance with the 125-word dictionary (Set A) augmented with garbage labels for 6 different garbage words and tested with tokens in vocabulary, out-of-vocabulary and garbage data sets

6 CONCLUSION

A robust speech detection algorithm for hands-free applications has been proposed which uses three technologies: noise suppression with efficient implementation, robust endpoint detection for non-stationary noise and speech verification with garbage modeling and confidence measure. Experiments show that high performance was obtained with the proposed method. From the results, it can be concluded that the proposed method can be applied for speech-activated hands-free systems such as cellular telephones, car navigation systems and entertainment robots.

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 Wu, D., M. Tanaka, R. Chen, L. Olorenshaw, , M. Amador and X. Menendez-Pidal "A Robust Speech Detection Algorithm for Speech Activated Hands-free Applications", ICASSP-99, pp. 2407-2410, Phoenix, April, 1999.

3 Caminero, J. et al, "On-line Garbage Modeling with Discriminant Analysis for Utterance Verification" Proceedings of ICSLP-96.

4 Boite J. M. et al, "A New Approach towards Keyword Spotting" Proceedings of Erospeech-93.

5 Duanpei Wu, X. Menendez-Pidal, L. Olorenshaw, R. Chen, M. Tanaka and M. Amador, "Speech and Word Detection Algorithms for Hands-Free Applications" Proceedings-97 of Sony Research Forum, Tokyo, 1999.