4

Voice and Noise Detection with AdaBoost

T. Takiguchi, N. Miyake, H. Matsuda, and Y. Ariki
Kobe University
Japan

1. Introduction

Speech recognition is one of our most effective communication tools when it comes to a hands-free (human-machine) interface. Most current speech recognition systems are capable of achieving good performance in clean acoustic environments. However, these systems require the user to turn the microphone on/off to capture voices only. Also, in hands-free environments, degradation in speech recognition performance increases significantly because the speech signal may be corrupted by a wide variety of sources, including background noise and reverberation.

Sudden and short-period noises also affect the performance of a speech recognition system. Figure 1 shows a speech wave overlapped by a sudden noise (a telephone call). To recognize the speech data correctly, noise reduction or model adaptation to the sudden noise is required. However, it is difficult to remove such noises because we do not know where the noise overlapped and what the noise was. Many studies have been conducted on non-stationary noise reduction in a single channel (A. Betkowska, et al., 2006), (V. Barreaud, et al., 2003), (M. Fujimoto & S. Nakamura, 2005). But it is difficult for these methods to track sudden noises.

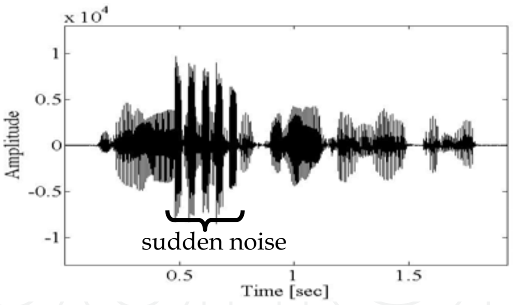


Figure 1. Speech wave overlapped by a sudden noise (telephone call)

In actual noisy environments, a speech detection algorithm plays an especially important role in noise reduction, speech recognition, and so on. In this chapter, a noise detection and classification algorithm using AdaBoost, which can achieve extremely high detection rates, is described. If a speech recognition system can detect sudden noises, it will make the system able to require the same utterance from the user again, and if clean speech data can

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

be input, it will help prevent system operation errors. Also, if it can be determined what noise is overlapped, the noise characteristics information will be useful in noise reduction.

"Boosting" is a technique in which a set of weak classifiers is combined to form one high-performance prediction rule, and AdaBoost (Y. Freund & R. E. Schapire, 1999) serves as an adaptive boosting algorithm in which the rule for combining the weak classifiers adapts to the problem and is able to yield extremely efficient classifiers. In this chapter, we discuss the AdaBoost algorithm for sudden-noise detection and classification problems. The proposed method shows an improved speech detection rate, compared to that of conventional detectors based on the GMM (Gaussian Mixture Model).

2. System Overview

Figure 2 shows the overview of the noise detection and classification system based on AdaBoost. The speech waveform is split into a small segment by a window function. Each segment is converted to the linear spectral domain by applying the discrete Fourier transform. Then the logarithm is applied to the linear power spectrum, and the feature vector (log-mel spectrum) is obtained. Next the system identifies whether or not the feature vector is a noisy speech overlapped by sudden noises using two-class AdaBoost, where the multi-class AdaBoost is not used due to the computation cost. Then the system clarifies the kind of sudden noises from only the detected noisy frame using multi-class AdaBoost.

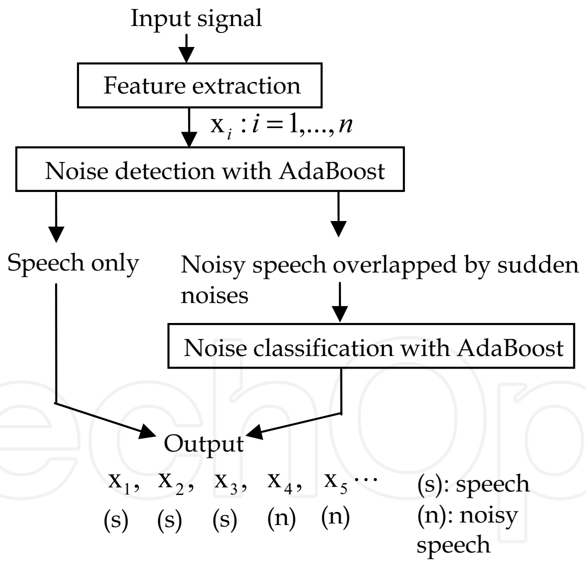

Input signal
Feature extraction
$x_i : i = 1,\ldots,n$
Noise detection with AdaBoost
Speech only
Noisy speech overlapped by sudden noises
Noise classification with AdaBoost
Output
$x_1,\ x_2,\ x_3,\ x_4,\ x_5 \cdots$
(s) (s) (s) (n) (n)
(s): speech
(n): noisy speech

Figure 2. System overview of noise detection and classification

3. Noise Detection with AdaBoost

Boosting is a voting method by weighted weak classifier and AdaBoost is one of method of Boosting (Y. Freund & et al., 1997). The Boosting decides the weak classifiers and their weights based on the minimizing of loss function in a two-class problem. Since the Boosting is fast and has high performance, it is commonly used for face detection in images (P. Viola, et al., 2001).

```
Input: $n$ examples $Z = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$
Initialize: $w_1(\mathrm{z}_i) = 1/n$ for all $i$ = 1,...,n
Do for $t$ = 1,...,$T$
    1. Train a base learner with respect to weighted example distribution $w_t$
       and obtain hypothesis $h_t : \mathrm{x} \mapsto \{\text{-}1,1\}$.
    2. Calculate the training error $\varepsilon_t$ of $h_t$.
       $\varepsilon_t = \sum_{i=1}^{n} w_t(\mathrm{z}_i) \frac{I(h_t(\mathrm{x}_i) \neq y_i) + 1}{2}$
    3. Set
       $\alpha_t = \log \frac{1 - \varepsilon_t}{\varepsilon_t}$
    4. Update example distribution.
       $w_{t+1} = \frac{w_t(\mathrm{z}_i) \exp\{\alpha_t \cdot I(h_t(\mathrm{x}_i) \neq y_i)\}}{\sum_{j=1}^{n} w_t(\mathrm{z}_i) \exp\{\alpha_t \cdot I(h_t(\mathrm{x}_i) \neq y_i)\}}$    (1)
Output: final hypothesis:
       $f(\mathrm{x}) = \frac{1}{|\alpha|} \sum_t \alpha_t \cdot h_t(\mathrm{x})$    (2)
```

Figure 3. AdaBoost algorithm for noise detection

Figure 3 shows the AdaBoost learning algorithm. The AdaBoost algorithm uses a set of training data, $\{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$, where x_i is the *i*-th feature vector of the observed signal and *y* is a set of possible labels. For the noise detection, we consider just two possible labels, $\mathrm{Y} = \{\text{-}1,1\}$, where the label, -1, means noisy speech, and the label, 1, means speech only.

As shown in Figure 3, the weak learner generates a hypothesis $h_t : \mathrm{x} \rightarrow \{\text{-}1,1\}$ that has a small error. In this paper, single-level decision trees (also known as decision stamps) are used as the base classifiers.

$$h_t(\mathrm{x}_i) = \begin{cases} 1, & \text{if } x_j \leq \theta_t \\ -1, & \text{else} \end{cases} \tag{3}$$

Here x_j is the *j*-dimensional feature of x and θ_t is the threshold which is decided by minimizing the error. After training the weak learner on the *t*-th iteration, the error of h_t is calculated.

Next, AdaBoost sets a parameter α_t. Intuitively, α_t measures the importance that is assigned to h_t. Then the weight w_t is updated. Equation (1) leads to the increase of the weight for the data misclassified by h_t. Therefore, the weight tends to concentrate on ``hard'' data. After the *T*-th iteration, the final hypothesis, $f(\mathrm{x})$ combines the outputs of the *T* weak hypotheses using a weighted majority vote. If $f(\mathrm{x}_i) < \eta$, AdaBoost outputs the label -1 and that means the *i*-th frame is a noisy frame overlapped by sudden noises to detect. In this paper, we set η = 0. As AdaBoost trains the weight, focusing on ``hard'' data, we can expect that it will achieve extremely high detection rates even if the power of noise to detect is low.

4. Noise Classification with Multi-Class AdaBoost

As AdaBoost is based on a two-class classifier, it is difficult to classify multi-class noises. Therefore, we use an extended multi-class AdaBoost to classify sudden noises. There are some ways to classify multi-class using a pair-wise method (such as a tree), K-pair-wise, or one-vs-rest (E. Alpaydin, 2004). In this paper, we used one-vs-rest for multi-class classification with AdaBoost. The multi-class AdaBoost algorithm is as follows:

Input: *m* examples $\{(\mathrm{x}_1, y_1), ..., (\mathrm{x}_m, y_m)\}$

$$y_i = \{1, ..., K\}$$

Do for *k* = 1,…,*K*

1. Set labels

$$y_i^k = \begin{cases} 1, & \text{if } y_i = k \\ -1, & \text{else} \end{cases} \tag{4}$$

2. Learn *k*-th classifier $f^k(\mathrm{x})$ using AdaBoost for data set

$$Z^k = (\mathrm{x}_1, y_1^k), ..., (\mathrm{x}_m, y_m^k)$$

Final classifier:

$$\hat{k} = \arg\max_k f^k(\mathrm{x}) \tag{5}$$

The multi-class algorithm is applied to the detected noisy frames overlapped by sudden noises. The number of classifiers, *K*, corresponds to the noise class. The *k*-th classifier is designed to separate the class *k* and other classes (Fig. 4) using AdaBoost described in Section 3. The final classifier decides a noise class having the maximum value from all classes in (5).

The multi-class AdaBoost can be applied to the noise detection problem, too. But in this paper, due to the computation cost, the two-class AdaBoost first detects noisy speech and then the detected frame only is classified into each noise class by multi-class AdaBoost.

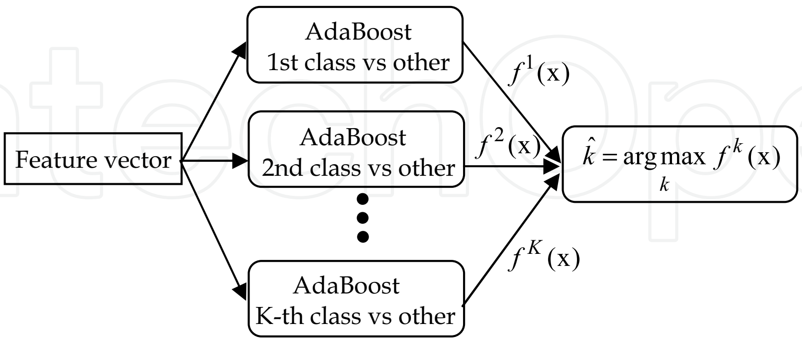


Figure 4. One-vs-rest AdaBoost for noise classification

5. GMM-Based Noise Detection and Classification

We used a conventional GMM (Gaussian mixture model) for comparing the proposed method. GMM is used widely for VAD (Voice Activity Detection) because the model is easy to train and usually powerful (A. Lee et al., 2004). In this paper, in order to detect sudden noises, we trained two GMMs, a clean speech model and a noisy speech model, where the number of mixtures is 64. Using two GMMs, the likelihood ratio is calculated by

$$L(\mathrm{x}) = \frac{\Pr(\mathrm{x} \mid \mathrm{speech\_model})}{\Pr(\mathrm{x} \mid \mathrm{noisy\_model})} \tag{6}$$

If $L(\mathrm{x}) > \theta$, **x** is detected as speech only. Otherwise **x** is detected as noisy speech.

In order to classify noise types, we need to train a noise GMM for each noise. Then, for the detected noisy speech only, we find a maximum likelihood noise from noise GMMs.

$$C(\mathrm{x}) = \arg\max_k \Pr(\mathrm{x} \mid \mathrm{noisy\_model}^{(k)}) \tag{7}$$

6. Experiments

6.1 Experimental Conditions

To evaluate the proposed method, we used six kinds of sudden noises from the RWCP corpus (S. Nakamura, et al., 2000). The following sudden noise sounds were used: spraying, telephone sounds, tearing up paper, particle-scattering, bell-ringing and horn blowing. Figure 5 shows the log-power spectrum of noises. In the database, each kind of noise has 50 data samples, which are divided into 20 data samples for training and 30 data for testing.

In order to make noisy speech corrupted by sudden noises, we added the sudden noises to clean speech in the wave domain and used 2,104 utterances of 5 men for testing and 210 utterances of 21 men for training (the total number of training data: 210 utterances × (6 + 1) = 1,470). The speech signal was sampled at 16 kHz and windowed with a 20-msec Hamming window every 10 msec, and 24-order log-mel power spectrum and 12-order

MFCCs were used as feature vectors. The number of the training iterations, T, is 500, where AdaBoost is composed of 500 weak classifiers.

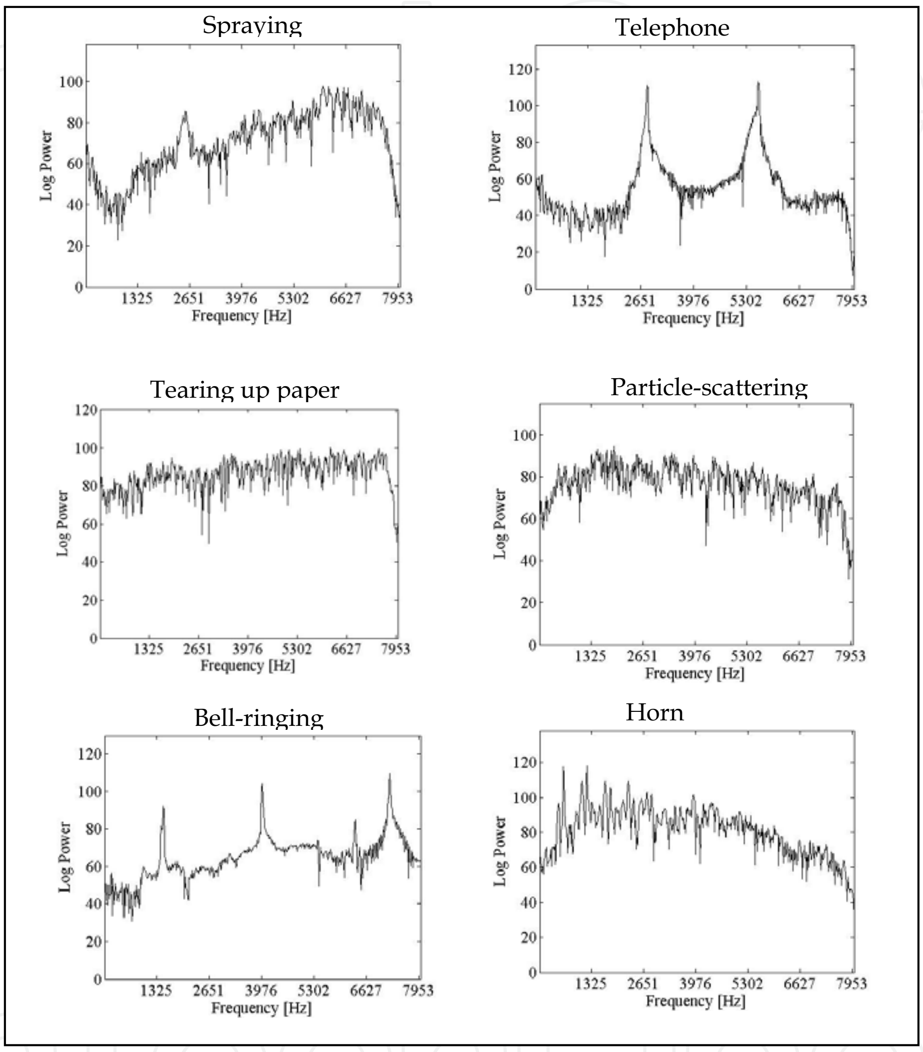


Figure 5. Log-power spectrum of noises

6.2 Experimental Results

Figure 6 and 7 show the results of the sudden-noise detection (F-measure) and classification (accuracy) at SNRs (Signal to Noise Ratio) of -5 dB, 0 dB, 5 dB and 10 dB. Here the SNR is calculated by

$$SNR = 10\log\left(\frac{E[s^2]}{E[n^2]}\right) \tag{8}$$

where $E[s^2]$ is the expectation of the power of the clean speech signal. Therefore, an increase of the SNR degrades the performance of the noise detection and classification because the noise power decreases. The F-measure used for the noise detection is given by

$$F = \frac{2 \cdot R \cdot P}{R + P}. \tag{9}$$

Here, R is recall and P is precision.

As can be seen from those figures, these results clarify the effectiveness of the AdaBoost-based method in comparison to the GMM-based method. As the SNR increases (the noise power decreases), the difference in performance is large. As the GMM-based method calculates the mean and covariance of the training data only, it may be difficult to express a complex non-linear boundary between clean speech and noisy speech (overlapped by a low-power noise). On the other hand, the AdaBoost system can obtain good performance at an SNR of 5 dB because AdaBoost can make a non-linear boundary from the training data near the boundary.

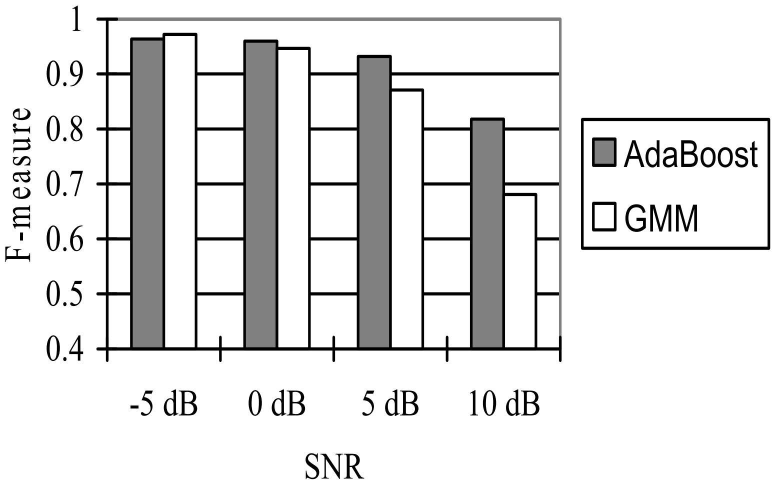


Figure 6. Results of noise detection

7. Conclusion

We proposed the sudden-noise detection and classification with Boosting. Experimental results show that the performance using AdaBoost is better than that of the conventional GMM-based method, especially at a high SNR (meaning, under low-power noise conditions). The reason is that Boosting could make a complex non-linear boundary fitting training data, while the GMM approach could not express the complex boundary because the GMM-based method calculates the mean and covariance of the training data only. Future research will include combining the noise detection and classification with noise reduction.

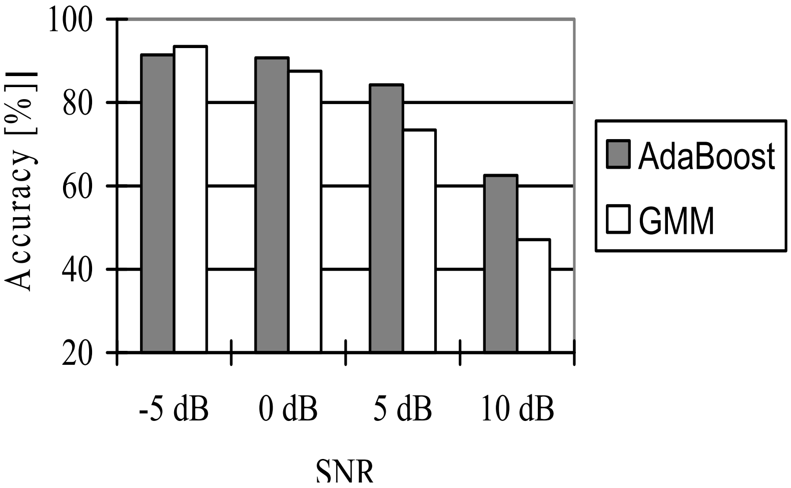


Figure 7. Results of noise classification

8. References

A. Betkowska, K. Shinoda & S. Furui (2006). FHMM for Robust Speech Recognition in Home Environment, *Proceedings of Symposium on Large-Scale Knowledge Resources,* pp. 129-132, 2006.

V. Barreaud, et al. (2003). On-Line Frame-Synchronous Compensation of Non-Stationary noise, *Proceedings of ICASSP*, pp. 652-655, 2003.

M. Fujimoto & S. Nakamura (2005). Particle Filter Based Non-stationary Noise Tracking for Robust Speech Recognition, *Proceedings of ICASSP*, pp. 257-260, 2005.

Y. Freund & R. E. Schapire (1999). A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, 14(5): pp. 771-780, 1999.

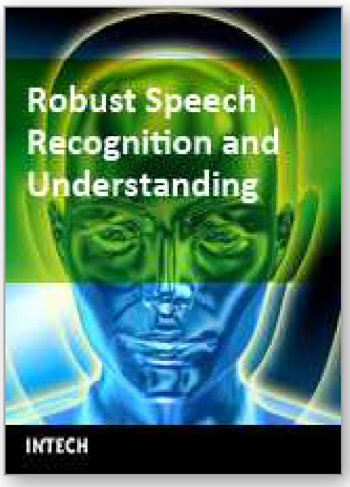
Y. Freund, et al. (1997). A decision-theoretic generalization of online learning and an application to boosting, *Journal of Comp. and System Sci.*, 55, pp. 119-139, 1997.

P. Viola, et al. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features, *IEEE CVPR,* vol. 1, pp. 511-518, 2001.

E. Alpaydin (2004). Introduction to Machine Learning, The MIT Press, ISBN-10: 0262012111.

A. Lee et al. (2004). Noise robust real world spoken dialog system using GMM based rejection of unintended inputs, *Proceedings of ICSLP*, pp. 173-176, 2004.

S. Nakamura et al. (2000). Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition, *Proceedings of 2nd ICLRE*, pp. 965-968, 2000.



This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.