©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Distributed audio network for speech enhancement in challenging noise backgrounds

Thorsten Kühnapfel, Tele Tan and Svetha Venkatesh
Dep. of Computing, Curtin University of Technology, Western Australia
(T.Kuehnapfel, T.Tan, S.Venkatesh)@curtin.edu.au

Burkhard Igel
Dep. of Information Tech. and Electrical Eng., University of Applied Sciences Dortmund, Germany
Igel@fh-dortmund.de

1. Abstract

This paper presents a new approach to enhance speech based on a distributed microphone network. Each microphone is used to simultaneously classify the input into either one of the noise types or as speech. For enhancing the speech signal a modified spectral subtraction approach is used that utilise the sound information of the entire network to update the noise model even during speech. This improves the reduction of the ambient noise, especially for non-stationary noise types such as street or beach noise. Experiments demonstrate the effectiveness of the proposed system.

2. Introduction

Observing large areas of interest is still a demanding challenge, whether it is done by CCTV, audio sensors or other motion sensors. Visual observation can provide many details of the scene but has proven very difficult to analysing the context. Therefore, audio sensors can provide vital information to enhance the understanding of the scene. Recording high quality speech signals in a real environment has its own challenges, due to ambient noise. The audio signal can be recorded either with one microphone or multiple microphones aligned in an array. Microphone arrays provide solutions for speech separation algorithms as beamforming or blind source separation. However, the main drawback when dealing with a large area of interest is that a sound signal, such as speech, decreases in signal intensity when the distance to the source increases. Therefore, it becomes costly to cover such an area with several microphone arrays. This paper proposes a system based on a network of single microphones, distributed around the area of interest, to record the audio signal and utilises the closest microphone to enhance speech signals.

A common approach to enhance speech for a single microphone is spectral subtraction. Generally, the background noise is modelled and is subtracted from the input signal. Several techniques have been proposed to model the noise: heuristically as in [2, 18] or statistically as in [4]. Depending on the accuracy of the model or the chosen parameters, spectral subtraction can introduce unwanted artefacts, known as "musical noise". The authors in [2] attempt to minimize the effect by improving the filter coefficients. The work in [18] focuses on a low resolution gain function. In [9] we present an approach to solve the problem of rapidly changing noise backgrounds during speech sequences by using multiple noise models.

Critical to any spectral subtraction algorithm is the voice activity detection (VAD). Early implementations are mainly based on the signal to noise ratio (SNR) as in [12], with the disadvantage of high false detection rates, especially for non-stationary noise [17]. More recent work [3] proposes to decompose the signal into sub-bands via perceptual wavelet-packet transformation and utilises masks to select critical sub-bands for speech. More sophisticated approaches based on statistics with smoothing have been proposed in [15]. This, however, is more complex and computationally intensive.

For any classification task, it is critical to choose a good set of features that characterise the data well. For audio data, many time and frequency domain features have been developed. In feature selection, information-theoretic approaches (e.g. Information Gain (IG) [16] or decision trees [14]) are needed to find the most discriminative features. Alternatively, the signal can be projected into a lower dimensional sub-space. A widely used linear transformation is principal component analysis (PCA) [8]. The authors in [11] present a PCA based approach to select audio features

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.17

IEEE computer society

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.

to detect bearing defects. Other work like [7] maps the features into the new reduced sub-space and uses the sub-space projection for classification.

The aim of the proposed system is to detect and enhance speech sequences in an environment with different, non-stationary types of ambient noise. For observing larger areas of interest, it becomes more economical to use a network of single microphones. Therefore, each audio stream is used to classify the noise by projecting the extracted audio features into a sub-space for each known noise source via PCA with the Mahalanobis distance used as distance metric. An advantage of this classification is that it can be used for detecting speech. For the proposed system the speech is estimated based on the classification result and the signal power estimation. The final overall noise classification is computed based on the entire network for all microphones that do not contain any speech. The audio stream of microphones with detected speech are then enhanced by a modified spectral subtraction approach [9], wherein different noise spectra are modelled to compensate for the changing ambient conditions. This spectral subtraction method also utilise the entire network for updating the estimated noise model during detected speech sequences.

The novelty of the proposed approach is that the result of the noise classification is used in combination with the signal power estimation to detect speech sequences in non-stationary ambient noise. Also the spectral subtraction approach is modified by updating the noise model during speech sequences based on the entire network. That enables the approach to compensate for the variance in the noise spectra for non-stationary noise types and improves the noise reduction during speech.

3. Methodology

This section details the components of the proposed surveillance system as shown in figure 1. Background noise classification is done by projecting the audio features into the sub-space and computing the Mahalanobis distance of the projected points to the projected cluster points of known noise models. Voice activity detection uses two observations: signal power and the estimated likelihood of the noise classification result. The VAD is also used to give a feedback to the signal power estimation in that during speech sequences, the mean signal power is not updated. For the final speech enhancement, spectral subtraction is applied to subtract the background noise.

3.1. Noise classification

Noise classification is performed in two stages: training where the known noise sources are modelled and subsequent classification. For both stages, the audio signal $y(i)$ is transformed into a feature set $f(i)$, where i is the time index. The features set $f(i)$ consist of the following 32 normalised features [13]: 13 Mel Frequency Cepstral Coefficient (MFCC), 15 energy sub-bands, zero crossing rate, spectral centroid, spectral spread, spectral skewness.

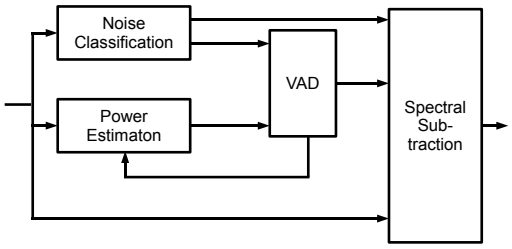

Figure 1. Signal flow of the proposed system

During training, PCA is used to reduce the dimensionality for each noise source n. First the feature set f^n is extracted for each noise source n and PCA is applied to derive the reduced sub-space, characterised by eigenvectors V^n. For each sub-space the same numbers of eigenvectors are selected. V^n can be used to transform the features f^n into the sub-space of noise source n as:

$$f'^n = (V^n)^T f^n \tag{1}$$

f'^n is used to compute the covariance Σ^n and mean μ^n of the noise source.

For classification, the original features are projected into the sub-space of each noise model and the Mahalanobis distance d [10] is computed as:

$$d^n(i) = \sqrt{(f'^n(i) - \mu^n)^T \Sigma^{n^{-1}} (f'^n(i) - \mu^n)} \tag{2}$$

where Σ^n and μ^n are the covariance and mean of the trainings samples of noise source n. The classification decision r^* is based on the shortest distance measure d over all noise types as:

$$r^*(i) = \underset{n}{\operatorname{argmin}}\, d^n(i) \tag{3}$$

The overall noise classification is based on a majority voting process of the entire microphone network. In general, let l^n be the count of classified noise type n over the entire network for each microphone where no speech is detected. The final noise classification r is computed as:

$$r(i) = \underset{n}{\operatorname{argmax}}\, l^n(i) \tag{4}$$

3.2. Voice activity detection

Voice activity detection is based on two measurements: the signal power of frequency sub-bands and the noise distance measure d^n. This combination provides reliable speech detection even with non-stationary noise sources wherein VAD based on signal intensity generally fails [17].

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.

3.2.1 VAD based on signal power estimation

The input signal $y(i)$ is transformed into the frequency domain via the FFT. Because the fundamental frequencies of the human voice for adult male and female ranges between 60 and 280 Hz [1], the speech signal has a bigger influence on the lower frequency range. Therefore for VAD, the proposed system uses 6 frequency sub-bands $P^k(i)$ ranging from 50 to 500 Hz, separated into linear sub-bands where k is the sub-band index. For each sub-band, the average signal power is estimated as:

$$\overline{P}^k(i) = (1-\gamma)\overline{P}^k(i-1) + \gamma P^k(i) \tag{5}$$

where γ is the smoothing factor which is set to zero when speech sequences are detected. This ensures that only the signal energy of the ambient noise is estimated.

Based on the signal intensity, voice activity is first detected for each sub-band individually, T^k, and then combined as R_P given by:

$$T^k(i) = \begin{cases} 1 \text{ ,if } P^k(i) \geq \psi \overline{P}^k(i-1) \\ 0 \text{ ,if } P^k(i) < \psi \overline{P}^k(i-1) \end{cases} \tag{6}$$

$$R_P(i) = \begin{cases} 1 \text{ ,if } \sum_k a_k T^k(i) \geq 0.5 \\ 0 \text{ ,if } \sum_k a_k T^k(i) < 0.5 \end{cases} \tag{7}$$

where a_k is a weighting factor and ψ is an empirical threshold that is defined in experiment section 4.2. a_k ensures that frequencies have a higher weighting between 60 to 280 Hz, with $\sum_k a_k = 1$. Speech is detected when $R_P = 1$. For a low signal to noise (SNR) ratio, ψ can be set lower which makes the VAD more sensitive but this will increase the probability of false detection.

3.2.2 VAD based on noise distance measure

The general idea of detecting speech sequences is that if speech is present, the distance d^n of the current noise source n is larger than with no speech. An unknown noise source would also have the same effect but can be adjusted by updating the known noise sources. To detect such a variance in d^n, a mean $\overline{d}^n$ is computed during training. The distance measure d of the classified noise source r (see equation 4) is used to check for presence of speech content as:

$$R_N(i) = \begin{cases} 1 & \text{,if } d^r(i) \geq \vartheta \overline{d}^r \\ 0 & \text{,if } d^r(i) < \vartheta \overline{d}^r \end{cases} \tag{8}$$

where ϑ is an empirical threshold and $R_N = 1$ indicates speech.

3.2.3 Combined VAD

The final detection of speech sequences is computed based on the signal power result R_P. The signal power measurement is chosen because it reliably indicates any change in signal intensity. To reduce the false detection, especially when non-stationary background noise is present, sequences must have certain elements of speech instances based on R_N. In our experiments, we chose sequences with at least 40% speech. A reason for this value is that in lower SNR situations, the distance d^n of the classified noise source is closer to the mean distance $\overline{d}^n$; because if the SNR decreases, the noise will further mask out the speech. That results in a better noise classification and lower speech detection rate.

3.3. Speech enhancement

Speech enhancement is achieved by spectral subtraction as proposed in [9]. In general, the recorded audio signal y is a mixture of the clean speech signal x and the background noise ω. A time varying filter with gain function G can be applied to the short-term frequency domain of y to estimate the speech signal $\hat{X}$ as:

$$\hat{X}(i) = G(i)Y(i) \tag{9}$$

and G is defined as:

$$G(i) = \max\left\{\sqrt{1-\alpha\frac{P_\omega(i)}{P_Y(i)}}, \beta\right\} \tag{10}$$

where P_ω and P_Y are the magnitude spectra of the modelled noise and input signal respectively, α is the subtraction factor and β is the floor function. P_ω must be updated to achieve the best possible speech estimation, especially for non-stationary noise sources. This should be done during noise sequences only with good results for quasi stationary noise, such as babble noise. If the noise is non-stationary, such as in street noise with passing cars, the estimation error of the modelled noise P_ω increases quickly, especially during speech sequences. To deal with this problem, we explore the use of the entire network to update the noise model during speech sequences, using the signal detection results of the microphones with no detected speech. For finding a microphone with similar sound characteristics, a correlation matrix of noise only sequences is computed. This correlation matrix is based on the correlation coefficient of the sound features f'^r over the time span of 1 second to account for the propagation time of sound. The noise model therefore is updated as:

$$P_\omega^m(i) = \begin{cases} (1-\eta)P_\omega^m(i-1) + \eta P_Y^c(i) & \text{,if speech} \\ (1-\eta)P_\omega^m(i-1) + \eta P_Y^m(i) & \text{,otherwise} \end{cases} \tag{11}$$

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.

where η is the smoothing factor, m is the current microphone index and c is the microphone closes to m with no detected speech.

4. Experiments

For all experiments the audio signal was sampled at 16 kHz with 16 bits per sample. The noises were recorded in real environments and either synthetically mixed with the speech signal using Audacity [19] or played back during sound recording in the anechoic chamber at the West Australian Telecommunications Research Institute (WATRI).

4.1. Noise classification

This experiment demonstrates the result of the noise classification of a single microphone when no speech is present. The sequential noise patterns introduced here are scooter, cafe, street and beach noise. Figure 2 shows the distance measure d^n of the projected test signal into each sub-space V^n.

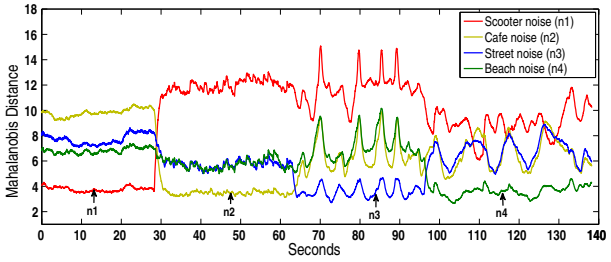

Figure 2. Mahalanobis distance for noise classification. Scooter noise is present from 0s till 28.8s, cafe noise till 63.4s, street noise till 96.7s and beach noise till the end.

The result shows that the distance measure of scooter and cafe noise has quite a low variance compared to street and beach noise. This is expected because scooter and cafe noise is more stationary than cars driving by or incoming waves. The classification rate for scooter, cafe, street and beach noise is 99.5%, 99.3%, 99.2% and 100% respectively. Only during noise changes some errors occur due to smoothing effects.

For V, we empirically selected the first 12 vectors based on the highest class separation and the overall noise classification accuracy. Figure 3 confirms this finding, showing that the first eigenvectors are the most important. After the 12th eigenvector the information gain is not as significant.

4.2. Speech parameter

This experiment evaluates the parameters range of ψ and ϑ for the voice activity detection. A speech signal of about 4s is digitally added to each noise sequence. It is fundamental to set the desired SNR value for the speech sequence, to ensure that further speech sequences with the same or a

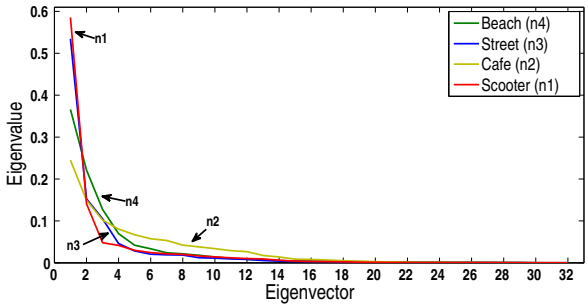

Figure 3. Eigenvalue for each eigenvector for scooter noise *n1*, cafe noise *n2*, street noise *n3* and beach noise *n4*.

greater SNR value are detected. For this experiment an average SNR value of -4.65dB, -4.26dB, 0.9dB and -2.02dB during speech activity is chosen for scooter, cafe, street and beach noise respectively. The values for beach and street noise is higher because the speech sequence is generally longer than a car passing by or an incoming wave. The minimum SNR value for beach and street noise are -5.4dB and -5.84dB respectively.

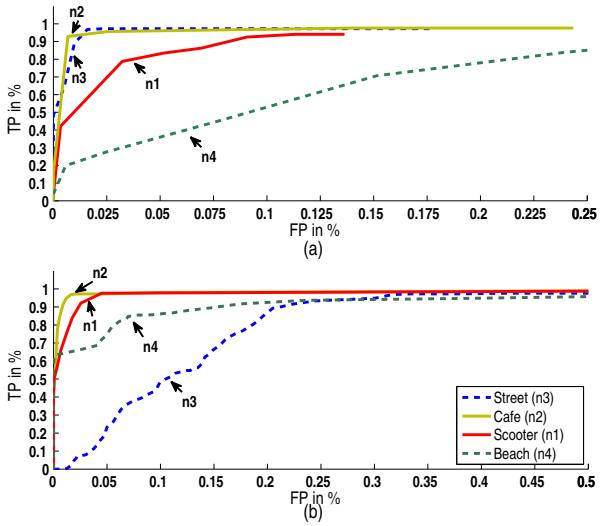

Figure 4. ROC graphs of the speech classification result based on varying ϑ and ψ respectively for R_N (a) and R_P (b). The true positive (TP) detection rate is represented along the y axis and the false positive (FP) along the x axis. Scooter noise is labelled as *n1*, cafe noise as *n2*, street noise as *n3* and beach noise as *n4*.

Figure 4 (a) shows that the speech sequence masked by beach noise is not that well detected when the decision is only based on the noise classification. The reason for this is that waves have a broad spectrum which is covering the speech signal. When it comes to VAD based on signal energy, the detection rate is better, however the false detection rate increases rapidly if a high accuracy is needed. Graph 4 (b) also shows that VAD for street noise based on signal intensity performs quite badly due to the rapid changing

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.

signal energy when cars are passing by. Based on these two graphs we chose ϑ as 1.2 and ψ as 1.5. The final speech detection result when both methods are combined is presented in table 1 (as described in section 3.2.3).

Noise	TP	FP
Scooter	97.24	3.16
Cafe	97.23	3.76
Street	98.03	4.56
Beach	82.65	10.53

Table 1. VAD result for combined approach measured in true (TP) and false (FP) positive detection rate, shown in %.

4.3. VAD

In this section, the proposed VAD is compared against the advanced front-end feature extraction algorithm (ES 202 050) [5] and the Support Vector Machine (SVM) [6]. The feature set for the SVM are 20 MFCC. Experiments involve two sequences where speech is masked by synthetic or real noise. The synthetic noise is white noise with a SNR of -6 dB and the other sequence contains scooter, cafe, street and beach noise with one speech sequence for each noise source as shown in figure 5.

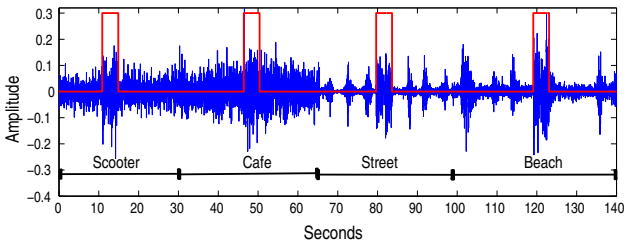

Figure 5. Test sequence with real noise types. Shown in red are the instances where speech is present.

Table 2 shows the result of the comparison for true positive and false positive classification results. For white noise all methods achieve a high true positive rate. Note that the proposed VAD produces the lowest false positive rate. For the real noise situation, the ES 202 050 is clearly unacceptable with an extremely high false positive rate. SVM detects only 42.2% of the speech sequences but compared to ES 202 050 has only a false positive rate of 9.3%. On the other hand, the proposed VAD is able to detect 98.7% of the speech sequences with a very low false detection rate of only 2.8%.

4.4. Speech enhancement

Speech enhancement is evaluate on the same audio file used in section 4.3 for real noise situation. Figure 6 shows the original signal (a) and a subsection of the enhanced

Method		White noise	Real noise
ES 202 050	TP	99.03	100
	FP	6.41	91.45
SVM	TP	97.94	42.23
	FP	8.15	9.31
Proposed VAD	TP	95.88	98.70
	FP	1.08	2.82

Table 2. VAD comparison for true positive (TP) and false positive (FP) classification results. All values are shown in %.

speech signal for the general spectral subtraction result [9] (b) and the proposed approach (c) as presented in section 3.3. It can be seen that (c) has a cleaner signal then (b).

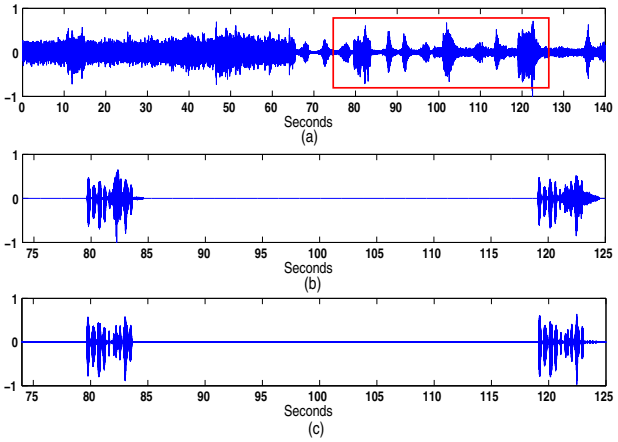

Figure 6. (a) shows the original audio signal and the subsection, marked by a rectangle, of the enhanced signal by the general spectral subtraction approach (b) and the proposed method (c).

To further verify the performances of these approaches, subjects were asked to rank the quality of the audio produced by these approaches. A mean opinion score (MOS) were then computed across these scores. The MOS is expressed as a single number ranging from 1 (Bad) to 5 (Excellent). The evaluation includes 2 aspects: the result of the ambient noise reduction and the quality of the enhanced speech.

Method	Score			
	Scooter	Cafe	Street	Beach
General [9]	4.45	3.95	2.86	3.09
Proposed	4.15	4.14	4.50	4.18

Table 3. Mean opinion score for ambient noise reduction for the general and proposed spectral subtraction approach.

Table 3 shows that the proposed subtraction method is able to remove more of the background noise than the general approach. For street and beach noise the difference is

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.

greatest because these noise types are highly non-stationary. The average score for the proposed method is 4.24 and for the general method is 3.59.

Method	Score			
	Scooter	Cafe	Street	Beach
General [9]	3.09	1.86	3.00	2.66
Proposed	3.23	2.95	3.77	3.73

Table 4. Mean opinion score for speech enhanced signal for the general and proposed spectral subtraction approach.

The speech evaluation shown in table 4 shows that the proposed spectral subtraction approach outperforms the general spectral subtraction approach in the area of enhanced speech quality. Again the proposed method shows the biggest advantage for highly non-stationary noise as street and beach noise. The average score for the proposed method is 3.42 and for the general method is 2.65.

5. Conclusion

We have demonstrated that the proposed system can reliably classify multiple non-stationary ambient noise sources. The classification outcome is also used to detect speech sequences. The experiments have proven that the combined approach using both signal intensity and the noise classification result has comparable performance for synthetic noise but outperforms the other methods when it comes to non-stationary real noise conditions.

For enhancing the desired speech signal, we presented a spectral subtraction approach which utilises the entire network. This approach was able to suppress the ambient noise even under non-stationary noise sources.

6. Acknowledgement

We thank Prof. Sven Nordholm and Eric Ostlin for providing the audio facilities and their assistance with the equipments.

References

[1] R. J. Baken and R. G. Daniloff. Readings in clinical spectrography of speech. *Singular Publishing Group, Inc.*, 1991. 3

[2] C. Breithaupt, T. Gerkmann, and R. Martin. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *IEEE Signal Processing Letters*, 14:1036–1039, 2007. 1

[3] S. Chena, H. Wua, Y. Chang, and T. Truong. Robust voice activity detection using perceptual wavelet-packet transform and teager energy operator. *Pattern Recognition Letters*, 28:1327–1332, 2007. 1

[4] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32:1109–1121, December 1984. 1

[5] ES 202 050 V1.1.5: Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. *ETSI*, 2007. 5

[6] S. Gunn. Support vector machines for classification and regression. Technical report, Department of Electronics and Computer Science, University of Southampton, 1998. 5

[7] T. Izumitani, R. Mukai, and K. Kashino. A background music detection method based on robust feature extraction. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13–16, 2008. 2

[8] J. Kittler. *Pattern regognition theory and application*, chapter Feature selection methodes on the Karhunen-Loeve expansion, pages 61–74. Noordhoff international publishing, 1977. 1

[9] T. Kühnapfel, T. Tan, S. Venkatesh, S. E. Nordholm, and B. Igel. Adaptive speech enhancement with varying noise backgrounds. *IEEE International Conference on Pattern Recognition*, December 2008. 1, 2, 3, 5, 6

[10] R. D. Maesschalck, D. Jouan-Rimbaud, and D. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50:1–18, 2000. 2

[11] A. Malhi and R. X. Gao. Pca-based feature selection scheme for machine defect classification. *IEEE Transactions on Instrumentation and Measurement*, 53:1517–1525, 2004. 1

[12] R. Martin. An efficient algorithm to estimate the instantaneous snr of speech signals. *Third European Conference on Speech Communication and Technology*, pages 1093–1096, 1993. 1

[13] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *IRCAM*, 2004. 2

[14] P. Perner and C. Apteb. Empirical evaluation of feature subset selection based on a real-world data set. *Engineering Applications of Artificial Intelligence*, 17:285–288, 2004. 1

[15] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio. Statistical voice activity detection using a multiple observation likelihood ratio test. *IEEE Signal Processing Letters*, 12:689– 692, 2005. 1

[16] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948. 1

[17] B. Wu and K. Wang. Noise spectrum estimation with entropy-based vad in non-stationary environments. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E89A:479–485, 2006. 1, 2

[18] L.-P. Yang and Q.-J. Fu. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The Journal of the Acoustical Society of America*, 117:1001–1004, 2005. 1

[19] http://audacity.sourceforge.net. 4

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 24,2010 at 05:41:25 UTC from IEEE Xplore. Restrictions apply.