Distributed audio network for speech enhancement in challenging noise backgrounds

Projektleitung
Prof. Dr. Burkhard Igel

Kooperationspartner
Curtin University
of Technology, Perth,
Australien

Zeitraum
2006–2009

Förderung
Deutsche Forschungsgemeinschaft (DFG)

Kontakt
Prof. Dr. Burkhard Igel
Fachbereich
Informations- und
Elektrotechnik
Fachhochschule
Dortmund
Sonnenstraße 96
44139 Dortmund
Tel.: 0231 9112-357
E-mail: igel
@fh-dortmund.de

Abstract

The objective of the research was to investigate a new approach for providing data of multiple sensors for tracking moving objects The last common 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.

Thorsten Kuhnapfel, Tele Tan and Svetha Venkatesh, Dep. of Computing, Curtin University of Technology, Western Australia

Burkhard Igel, Dep. of Information Tech. and Electrical Eng., University of Applied Sciences Dortmund, Germany

Methodology

Each audio stream is used to classify the noise by projecting the extracted audio features into a subspace for each known noise source via principal component analysis (PCA) with the Mahalanobis distance [1] used as a distance metric. 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 (VAD) 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.

Voice activity detection

The developed algorithm is compared to the advanced front-end feature extraction algorithm (ES 202 050) [2] and the Support Vector Machine (SVM) [3]. 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 the following figure.

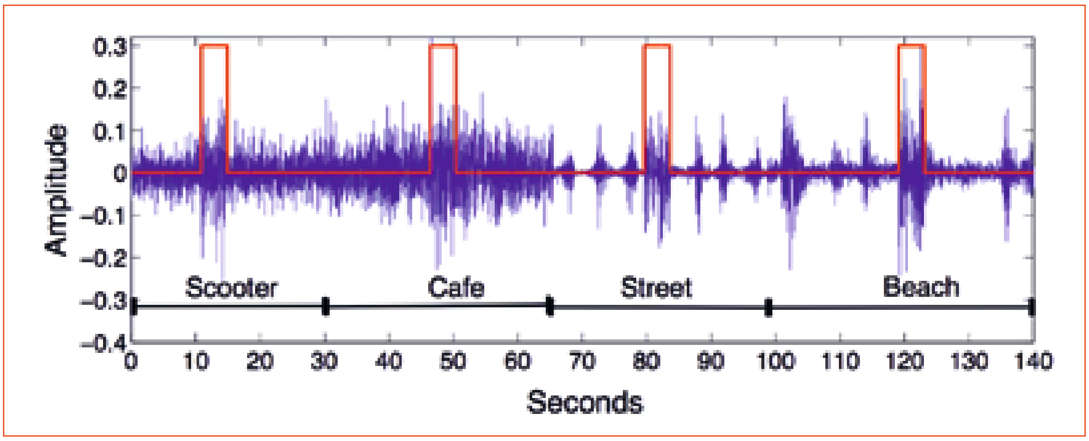

Figure 1:
The general idea of detecting speech sequences is that if speech is present, the distance d^n of the current noise source n is larger than with no speech. An unknown noise source would also have the same effect but can be adjusted by updating the known noise sources. To detect such a variance in d^n, a mean d^n is computed during training.

Speech enhancement is evaluate on this audio file for real noise situation. Figure 2 shows the original signal (a) and a subsection of the enhanced speech signal for the general spectral subtraction result [4] (b) and the proposed approach (c). It can be seen that (c) has a cleaner signal then (b).

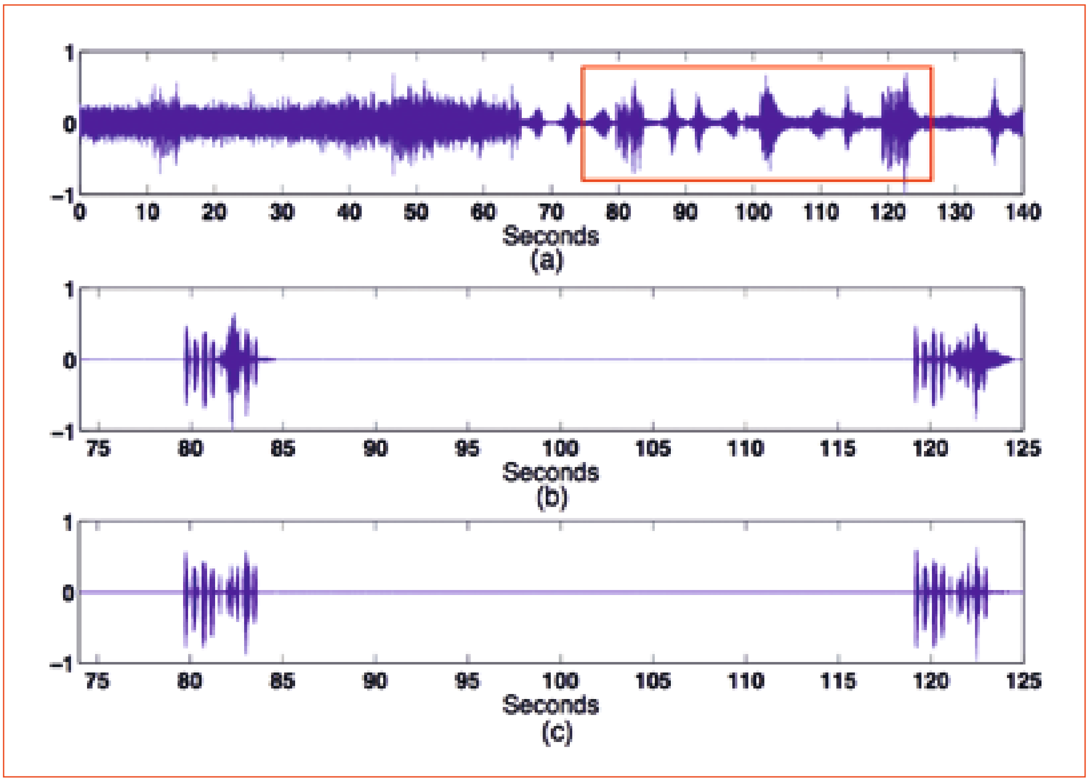

Figure 2:
Figure 2 (a) shows the original audio signal and the subsection, marked by a rectangle, of the enhanced signal by the general spectral subtraction approach (b) and the proposed method (c). To further verify the performances of these approaches, subjects were asked to rank the quality of the audio produced by these approaches. The speech evaluation shown shows that the proposed spectral subtraction approach outperforms the general spectral subtraction approach in the area of enhanced speech quality.

Conclusion

We can demonstrate 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.

Details:

See full paper accepted on AVSS 2009 - 6th IEEE International Conference on

Advanced Video and Signal Based Surveillance, Genoa, Italy, September 2-4, 2009

References:

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

[2] 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.

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

[4] T. Kuhnapfel, 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.