Robust Voice Activity Detection for Social Sensing

Sebastian Feese
Wearable Computing Lab.
ETH Zurich
feese@ife.ee.ethz.ch

Gerhard Tröster
Wearable Computing Lab.
ETH Zurich
troester@ife.ee.ethz.ch

UbiComp'13 Adjunct, September 8–12, 2013, Zurich, Switzerland.

http://dx.doi.org/10.1145/2494091.2497347

Abstract
The speech modality is a rich source of personal information. As such, speech detection is a fundamental function of many social sensing applications. Simply the amount of speech present in our surroundings can give indications about our socialbility and communication patterns. In this work, we present and evaluate a speech detection approach utilizing dictionary learning and sparse signal representation. Transforming the noisy audio data to the sparse representation with a dictionary learned from clean speech data, we show that speech and non speech can be discriminated even in low signal-to-noise conditions with up to 92% accuracy. In addition to an evaluation with simulated data, we evaluate the algorithm on a real-world data set recorded during firefighting missions. We show, that speech activity of firefighters can be detected with 85% accuracy when using a smartphone that was placed in the firefighting jacket.

Author Keywords
robust speech detection; human behavior observation; communication pattern; social sensing

ACM Classification Keywords
H.1.2 [User/Machine Systems]; H.5.5 [Sound and Music Computing]; J.4 [Social and Behavioral Sciences]

Introduction

Speech is an important modality which reveals personal information, e.g. about our state of mind, our emotions and connection to others. In groups of persons, communication patterns can indicate social structure and can characterize relationships. Within work teams, simply the amount of speech can indicate explicit coordination within teams.

Within the interdisciplinary SNSF-funded research project "Micro-level behavior and team performance", we apply social sensing to team research. One of our goals is to measure communication patterns in first responder teams such as firefighters automatically with the smartphone. Noisy work environments and the placement of the smartphone in the firefighting jacket require a robust voice activity detection in order to estimate the amount of communication accurately in the field. The fact that the detection system must work across various noise types and at different signal-to-noise levels renders the detection task challenging.

In order to detect speech in noisy ambient sound recorded with the smartphone we rely on dictionary learning and sparse representation. Our contributions are the following:

1. We present a noise robust voice activity detection system based on dictionary learning and sparse representation.
2. We evaluate the approach on simulated data using the TIMIT and NOIZEX-92 databases.
3. We test the voice activity detection algorithm on ambient sound data recorded with the smartphone during firefighting trainings.

Related Work

In recent years the smartphone became a true sensing platform and enabled ubiquitous sensing of human behavior. Previous research has shown how user context and behavior can be inferred from different sensor modalities. Particularly, ambient sound proved to be a rich source of personal information. The built-in microphone of smartphones was utilized to sense ambient sound patterns [6], to recognize emotions of the user [8], to detect user conversations [5], as well as to indicate levels of socialbility as one factor of well-being [4]. However, most these applications were designed for office use were noise conditions are at an acceptable level to make inferences about personal states. Outdoors, for example on noisy streets, already the detection of speech becomes challenging.

In the signal processing community various robust voice activity detectors have been developed which work in noisy environments. For example the long-term-spectral-variability (LTSV) introduced in [3] measures the non-stationarity of signals. Because speech and noise exhibit different levels of non-stationarity the measure can be used for voice activity detection. Recently an approach for robust activity detection based on dictionary learning and sparse representation was proposed in [10]. In this work, we compare the two approaches to detect speech / non speech in ambient sound recordings collected with the smartphone placed in a jacket pocket.

Noise Robust Speech Detection

Approach

Our approach to detect speech in noisy environments utilizes dictionary learning and sparse representation of the noisy audio signal. Using a dictionary learned on clean speech data, the sparse representation better

approximates speech than noise signals and thus can be used for speech detection. Because of the sparsity constraints, speech can be detected even in low signal-to-noise conditions when speech is barely audible.

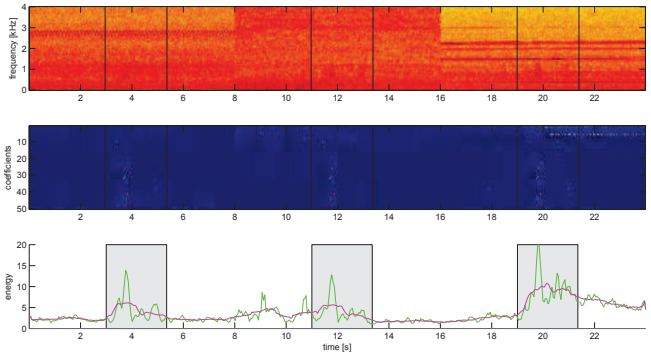

Figure 1: Sparse representation of noisy speech. Top: spectrogram of noisy signal; Middle: sparse coefficients; Bottom: total coefficient energy per frame smoothed with short and long term sliding windows.

The approach is illustrated in Figure 1. In the example, one sentence from the TIMIT database [2] was mixed with three different noise types from the NOIZEX-92 database [9] at -10 dB SNR. From the spectrogram, the difficulty of the detection at such a low noise level becomes apparent. However, in the sparse representation the detection task becomes feasible, as the squared coefficients highlight voiced parts of the spoken sentence. As can be seen, the total energy of the sparse coefficients is much higher for speech than for noise and peaks at the voiced parts of speech (bottom). Comparing short-term and long-term averages of the total coefficient energy has been shown to robustly detect speech [10].

Recognition Chain

The speech detection chain is presented in Figure 2. The audio signal is framed using a hamming window and then transformed into the sparse representation. Frames of the sparse representation are used to calculate features on longer windows which are fed into a classifier for speech / non speech detection.

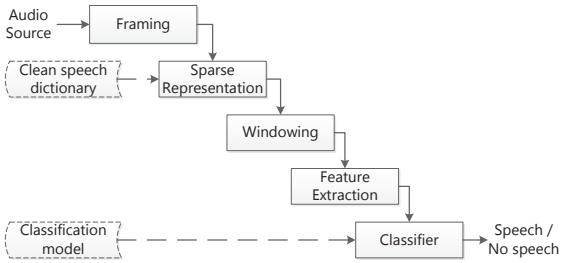

Figure 2: Proposed voice activity detection chain

Dictionary Learning

To learn a dictionary from data, one searches the dictionary $\boldsymbol{D}_{opt}$ that best represents the training data $\boldsymbol{X} = [\boldsymbol{x_1}, .., \boldsymbol{x_n}]$, while having a sparse solution. This is expressed by the l_1-sparse-coding problem:

$$\boldsymbol{D}_{opt} = \underset{\boldsymbol{D},\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (\|\boldsymbol{x_i} - \boldsymbol{D}\boldsymbol{\alpha_i}\|_2^2 - \lambda \|\boldsymbol{\alpha_i}\|_1), \quad (1)$$

where λ is the regularization parameter corresponding to the effective sparsity of the solution.

For learning the dictionary from clean speech data, we randomly sample frames $\boldsymbol{x_i}$ of length fl of 200 randomly selected sentences from the TIMIT database [2]. Because voiced parts of speech are most discriminative in noisy conditions, we only consider frames that include at least

80% of voiced speech. Each frame is multiplied by a hamming window. In total, we sample 10^6 frames for each considered frame length. We use the online method of Mairal et. al. [7] to solve Equation 1.

To illustrate the learned dictionary, we present in Figure 3 the spectrograms of each dictionary atom. As can be seen, the learned dictionary atoms appear to be similar to voiced parts of speech.

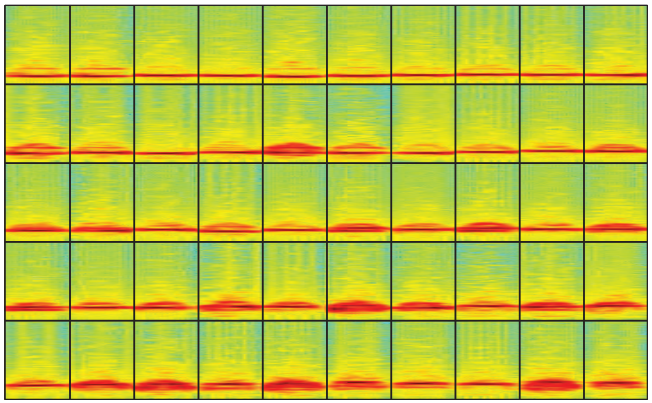

Figure 3: Learned Dictionary ($k = 50$, $fl = 100$ms): For each of the 50 atoms the spectrogram is presented.

Sparse Representation
To find a sparse representation $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, .., \boldsymbol{\alpha}_n]$ of the audio signal given by frames $[\boldsymbol{x_1}, .., \boldsymbol{x_n}]$ one needs to find the coefficients that minimise the representation error while being sparse. Similar to above this is expressed by:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (\|\boldsymbol{x_i} - \boldsymbol{D}_{opt}\boldsymbol{\alpha_i}\|_2^2 - \lambda\|\boldsymbol{\alpha_i}\|_1) \qquad (2)$$

Classification
The classification of speech / non speech is done on hopping windows of length W and step size S. Having observed that voiced speech has high energy coefficients in the sparse representation whereas noise signals have low energy coefficients (compare Figure 1), we compute the maximum total coefficient energy within a window and subtract the median value for reasons of normalisation. Normalisation is necessary due to different noise types. In formula, the feature for each window i is given by:

$$f(i) = \max_{i<=j<i+W} e(j) - \operatorname{median}_{i<=j<i+W} e(j), \qquad (3)$$

where $e(j)$ is the total coefficient energy of frame j by $e(j) = \|\boldsymbol{\alpha_j}\|_2^2$.

For classification we use logistic regression. All results reported below used 10-fold cross-validation.

Evaluation

For evaluation we randomly selected 96 sentences of the TIMIT database that were not previously used for the dictionary learning and concatenated them with 3 seconds of silence in between. The selected sentences had an average length of 3.25 ± 0.95 seconds. The clean speech data was mixed at three different SNRs (0,-5,-10 dB) with 12 noize types of the NOIZEX-92 database from which we did not include 'babble' and 'destroyerop' because they included speech. SNR was only calculated when speech was present. In total 360 minutes of audio data were used for the evaluation which was done independently of noise type and SNR.

Dictionary Parameters
To find suitable dictionary parameters, we compared the detection accuracy for different frame lengths and dictionary sizes. All other parameters were fixed: the

regularisation parameter λ was set to 0.15, the frame overlap to 50%, window length W to 1 second and step size S to 100 ms. In Figure 4 the accuracies of the different combinations are presented. It can be seen that the best detection performance on 1 second long prediction windows is reached at a frame length of $fl = 100$ms and a dictionary size of $k = 50$.

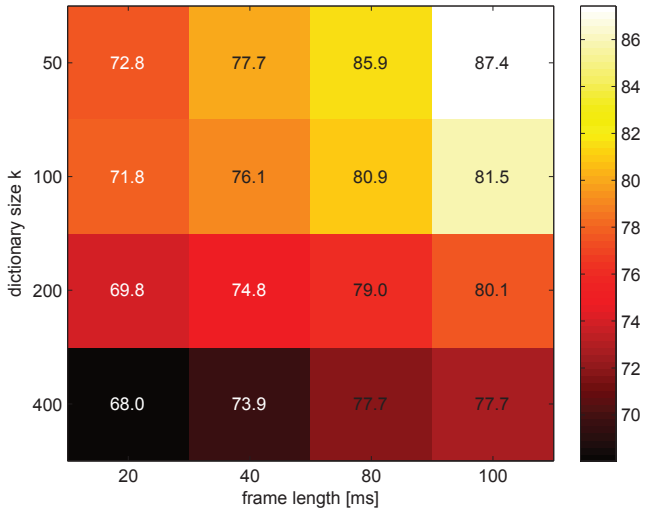

Figure 4: Average detection accuracies across different noise types and levels at different dictionary sizes and frame lengths.

Comparison to LTSV

We compared the presented approach to another robust voice activity detector based on the long-term-spectral-variability measure introduced in [3]. LTSV was calculated with parameters as presented in [3] on frames of 20 ms length and a step size of 10 ms. Similar to above, the frame based measure was aggregated on a longer window length. For each window the root-mean-square over all LTSV-frames included in one window was calculated and classified using logistic regression. The results are presented in Figure 5 for different window lengths. As can be seen, LTSV is better than the sparse representation approach (SR) for window lengths shorter 800 ms, whereas for longer windows SR is better. In both cases detection accuracy increase with longer windows. The fact that LTSV is better at shorter window sizes is due to the fact, that internally LTSV includes information of 1 second long windows.

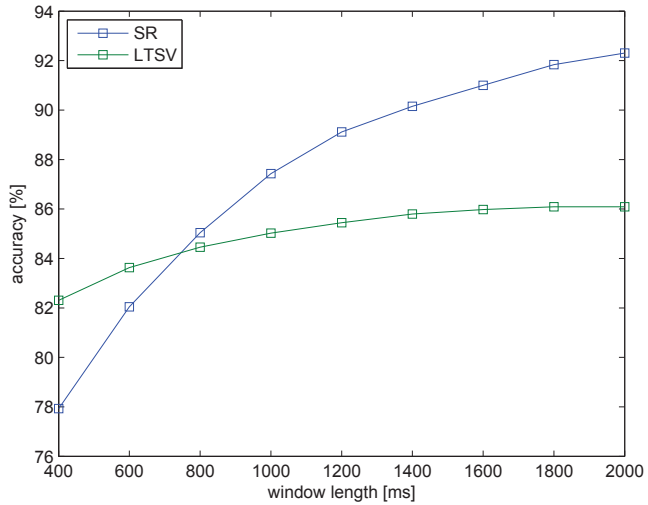

Figure 5: Comparison of average detection accuracies across different noise types and levels at different window lengths. $k = 50$, $fl = 100$ms, $S = 50$ms.

Speech Detection during Firefighting

In order to sense the amount of team communication within first responder teams, we have tested the speech

detection algorithm on audio data that was recorded during a one day training of firefighters.

Experiment
The experiment was conducted in the fire simulation building at the training facilities of the Zurich fire brigade where a variety of fire scenarios can be realistically simulated ranging from kitchen fires to a burning car in the garage. In the chosen scenario a kitchen fire in the third floor of the training building had to be extinguished.

Two teams of a voluntary fire brigade completed the scenario one after the other. Each team consisted of five firefighters including the incident commander (IC) who led mission operations and the troop leader (TL) who led the troop that went inside the building to extinguish the fire. To coordinate mission operations, incident commander and troop leader had to communicate. Impressions of the scenario are shown in Figure 6.

Figure 6: Smartphone placement and impressions of the firefighting training scenario.

For data collection, we used the Sony Xperia Active smartphone and a custom Android app. Based on the funf-open-sensing-framework[1], we designed an Android app to record ambient sound data at a sample rate of 11250 Hz and later down sampled to 8 kHz. The phone was placed in the left pocket of the firefighting jacket (see Figure 6) where firefighters were used to carry their mobile phone. For more details on the experiment please refer to [1].

Test on Firefighting Noise
In order to test the accuracy of the detection algorithms on noise types that are observed during firefighting, we manually selected eight different noise snippets from the ambient sound recorded during the training missions. This included engine noise of the fire truck, rustling noise when waking, background noise of the fire house such as a loud fan and breathing noise when using the self contained breathing apparatus. As above, these noise types were mixed with the same clean speech data at the three different noise levels. In Figure 7 the results on the simulated noisy speech data are presented. It can be seen, that the SR approach to speech detection is robust also to typical noise types observed in a firefighting training scenario. The average accuracy over all SNRs and noise types at a window length of 2 seconds is above 90%. Compared with the LTSV approach the detection accuracies are about 8% higher.

[1]http://funf.org/

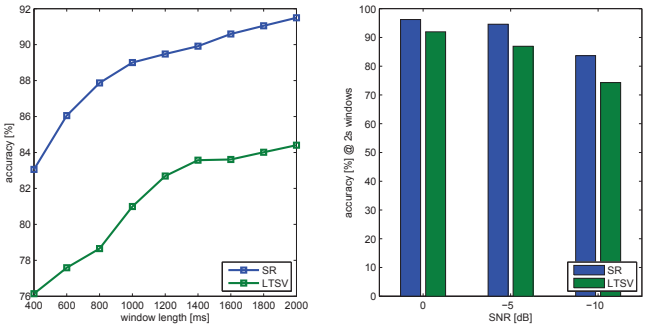

Figure 7: Voice activity detection accuracies when clean speech was mixed with typical noise types observed during a firefighting training mission.

Test on Firefighting Audio Data

To test our speech detection algorithm on noisy speech data observed during firefighting, we manually labeled 50 minutes of the recorded audio data for present speech at the incident commanders and troop leaders.

In the left of Figure 8 the accuracies are shown. As can be seen the voice activity detection works about 10% better for the incident commanders (IC1, IC2) who were outside the building compared to the troop leaders (TL1, TL2) who were inside. This difference in detection accuracy can be explained by different levels of environmental noise as the building ventilation was very noisy. In the right of Figure 8 the event-based F-measure is presented. At a window length of 1 second, the F-measure is above 85% for all four firefighters, meaning that only very few speech events were inserted or deleted.

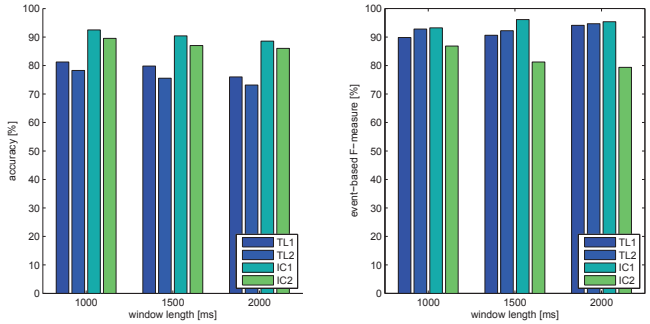

Figure 8: Continuous voice activity detection results for ambient sound data recorded during firefighting training. Left: accuracy; Right: event-based F-measure

Conclusion

We have presented a robust voice activity detection algorithm which is based on sparse representation. To best represent speech, we used a dictionary learned from clean speech data. The evaluation on simulated noisy speech data proofed robustness even in low signal-to-noise conditions. On average an accuracy of 87%, 92% was reached on a window length of one, two seconds, respectively. To test real-world noisy scenarios, we applied the detection algorithm to ambient sound data which was recorded during firefighting trainings. On average an accuracy of 85% and an event-based F-measure of 91% was obtained on a window length of 1 second. Future work should address the problem of speaker diarization in low signal-to-noise conditions.

Acknowledgements

This work is partly funded by the SNSF interdisciplinary project "Micro-level behavior and team performance" (grant agreement no.: CR12I1_137741).

References

[1] Feese, S., Arnrich, B., Rossi, M., Burtscher, M., Meyer, B., Jonas, K., and Tröster, G. Towards Monitoring Firefighting Teams with the Smartphone. In *Proc. WIP PerCom* (2013).

[2] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. TIMIT acoustic phonetic continuous speech corpus, 1993.

[3] Ghosh, P. K., Tsiartas, A., and Narayanan, S. Robust Voice Activity Detection Using Long-Term Signal Variability. *IEEE Trans. ASLP* (2011).

[4] Lane, N., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., and Campbell, A. BeWell: A Smartphone Application to Monitor, Model and Promote Wellbeing. In *Proc. PervasiveHealth* (2011).

[5] Lu, H., Brush, A. J. B., Priyantha, B., Karlson, A. K., and Liu, J. Speakersense: energy efficient unobtrusive speaker identification on mobile phones. In *Proc. Pervasive* (2011).

[6] Lu, H., Pan, W., Lane, N. D., Choudhury, T., and Campbell, A. T. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proc. MobiSys* (2009).

[7] Mairal, J., Bach, F., and Edu, G. U. M. N. Online Dictionary Learning for Sparse Coding. In *Proc. ICML* (2009).

[8] Rachuri, K. K., Mascolo, C., Rentfrow, P. J., and Longworth, C. EmotionSense : A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *Proc. UbiComp* (2010).

[9] Varga, A., and Steeneken, H. J. M. Assessment for automatic speech recognition ii: Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* (1993).

[10] You, D., Han, J., Zheng, G., and Zheng, T. Sparse power spectrum based robust voice activity detector. In *Proc. ICASSP* (2012).