An Integrated Solution for Snoring Sound Classification Using Bhattacharyya Distance based GMM Supervectors with SVM, Feature Selection with Random Forest and Spectrogram with CNN

Tin Lay Nwe, Huy Dat Tran, Wen Zheng Terence Ng, Bin Ma

Institute for Infocomm Research, I2R, Singapore

tlnma@i2r.a-star.edu.sg, hdtran@i2r.a-star.edu.sg, wztng@i2r.a-star.edu.sg, mabin@i2r.a-star.edu.sg

Abstract

Snoring is caused by the narrowing of the upper airway and it is excited by different locations within the upper airways. This irregularity could lead to the presence of Obstructive Sleep Apnea Syndrome (OSAS). Diagnosis of OSAS could therefore be made by snoring sound analysis. This paper proposes the novel method to automatically classify snoring sounds by their excitation locations for ComParE2017 challenge. We propose 3 sub-systems for classification. In the first system, we propose to integrate Bhattacharyya distance based Gaussian Mixture Model (GMM) supervectors to a set of static features provided by ComParE2017 challenge. The Bhattacharyya distance based GMM supervectors characterize the spectral dissimilarity measure among snore sounds excited by different locations. And, we employ Support Vector Machine (SVM) for classification. In the second system, we perform feature selection on static features provided by the challenge and conduct classification using Random Forest. In the third system, we extract spectrogram from audio and employ Convolutional Neural Network (CNN) for snore sound classification. Then, we fuse 3 sub-systems to produce final classification results. The experimental results show that the proposed system performs better than the challenge baseline.

Index Terms: snore sound classification, GMM supervectors, computational paralinguistics

1. Introduction

Snoring occurs when there is a narrowing or obstruction in the upper airways, the turbulence of air and the vibrations of soft palate [1]. Characteristics of the snore sound change depending on the intensity of the vibration on different locations (or excitation location) within the upper airway. Snoring is usually not serious and it is a common condition. However, irregularities in snoring may be associated with sleep apnea [2].

Over the past years, several studies [3], [4] have been done to detect or classify snore sound from other types of sounds such as breath, noise, etc.,. Several acoustic features are explored to detect snore sound. In [3], zero crossings, logarithm of the signal energy and the first formant are used as features to classify the sound segments into either snore or breath classes. In [5], several temporal and spectral components are explored to detect snoring sound from non-snoring sounds. Temporal components include duration, average energy, skewness and zero-crossing rate (ZCR). And, spectral features are average power and the first three formant frequencies for each segment. In [6], spectral-based features are used for detecting snoring sound. In [7], features such as energy, formant, Mel-scale frequency cepstral coefficients (MFCC) are used to detect snore related signals.

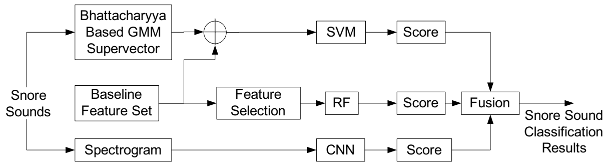

Figure 1: *System block diagram for classification of snore sounds based on their excitation location.*

The previous studies indicate that snore and non-snore sounds could be categorized automatically with a high degree of accuracy [4]. However, very few works have been done to classify the snore sounds based on their excitation location. A study which classifies snoring sound based on excitation location is classification between two types of snore sounds: oral and nasal snoring sounds [8]. This study uses fundamental frequency and the maximum of the amplitude spectrum in the specific band to classify two types of snoring sounds. However, there is still a shortage of robust methods to automatically detect snoring sounds based on their excitation location. The treatment will be effective if it is targeting the area in the upper airways where the snoring sound is generated. In this paper, we propose 3 sub-systems for classifying 4 types of snore sounds based on their excitation location. We then fuse these 3 sub-systems to generate final classification results.

In the first system, we integrate Bhattacharyya based GMM supervector to the baseline feature set which is provided by challenge. Bhattacharyya based GMM supervector characterizes Spectral Dissimilarity (SD) measure between the sounds. In the second system, we use Correlation Feature Selection (CFS) method to select the subset of features from challenge provided baseline feature set. In the third system, we explore local acoustic characteristics in time-frequency spectrogram to classify snore sounds. We employ machine learning methods such as Support Vector Machine(SVM), Random Forest (RF) and Convolutional Neural Network(CNN) as classifiers for the three sub-systems respectively. System block diagram of our snore sound classification is shown in Figure 1.

The rest of the paper is organized as follows. Section 2 presents the first sub-system that integrate Bhattacharyya distance based GMM supervectors that characterizes Spectral Dissimilarity(SD) measure to baseline feature set. Section 3 presents the second sub-system with feature selection and Random Forest classifier. Section 4 explains the third sub-system that uses spectrogram as input to CNN classifier. Section 5 describes database and baseline features. Section 6 presents experimental results and Section 7 concludes the paper.

2. Bhattacharyya based GMM supervectors and SVM Classifier

Four types of snore sounds considered correspond four different excitation locations within upper airways. And, there are dissimilarities in acoustic characteristics between different types of snore sounds. Hence, directly using the dissimilarity measure as the feature parameters, can better describe the essence of snore sounds. A standard GMM-supervector characterizes the Spectral Dissimilarity (SD) measure between sounds by the GMM parameters such as the mean vectors and covariance matrices. We employ GMM-SVM kernel with Bhattacharyya based GMM distance to measure SD between snore sounds. Beside the first-order statistics of mean, we consider SD measure using second-order statistics of covariance which describe the shape of the distribution. SD measure based on first-order statistics of mean gives the major characteristics of the probabilistic distance. We discuss the standard GMM supervector in the following section.

2.1. Standard GMM-supervector

The density function of a GMM is defined as in equation (1).

$$p\left(x\right) = \sum_{i=1}^{M} \omega_i f\left(x|m_i, \Sigma_i\right) \tag{1}$$

where $f\left(.\right)$ denotes the Gaussian density function. And, m_i, Σ_i and ω_i are the mean, covariance matrix and weight of i^{th} Gaussian component, respectively. M is number of Gaussian mixtures. And, x is a D-dimensional acoustic feature vector. The standard GMM-supervector is the stacked normalized mean vectors of the GMM.

2.2. GMM-supervector with Bhattacharyya based kernel

Bhattacharyya distance is a separability measure between two Gaussian distributions as in equation (2). In equation (2) Σ_i^a and Σ_i^b are the covariance matrices of the sounds adapted from Universal Background Model (UBM). And, m_a^i and m_b^i are the adapted mean vectors. Σ_i^u is the covariance matrix of the UBM. p_a and p_b are the probabilistic models, GMM_a and GMM_b, respectively. The first term of equation (2) gives the class separability due to the difference between class means, while the second term gives the class separability due to the variance between class covariance. Based on the first two terms, Bhattacharyya distance based kernel and Bhattacharyya distance based GMM supervector are formulated in [9]. The i^{th} subvector of Bhattacharyya distance based GMM-supervector for a snore sound is shown in equation (3) [9]. GMM-supervector with Bhattacharyya based kernel is obtained by stacking all i^{th} subvectors of equation (3).

$$g^{Bhat}(m_i, \Sigma_i) = \begin{bmatrix} \left(\frac{\Sigma_i^{\lambda}+\Sigma_i^{u}}{2}\right)^{-1/2} \left(m_i^{\lambda} - m_i^{u}\right) \\ \\ diag\left(\left(\frac{\Sigma_i^{\lambda}+\Sigma_i^{u}}{2}\right)^{1/2} \left(\Sigma_i^{\lambda}\right)^{-1/2}\right) \end{bmatrix} \tag{3}$$

Snore sounds of same category have similar spectral distributions while those of different categories have different distributions. We use the Spectral Dissimilarity (SD) measure in terms of spectral distributions together with challenge provided baseline feature set as features for SVM classifier.

If we look at equation (3), the first term reflects the dissimilarity between mean of a snore sound and that of a Universal Background Model (UBM). This mean statistical dissimilarity gives the major characteristics of the probabilistic distance. Besides the first-order statistics of mean, the second-order statistics of covariance matrices describing the shape of the spectral distribution is also useful to measure SD. If we look at the second term of equation (3), it represents the ratio between covariance of a UBM and that of a snore sound. In other word, this second term describes dissimilarity in shape of the spectral distribution. Hence, these two terms represent the SD measure between a snore sound from a reference UBM. As snore sounds with different excitation locations have different spectral distributions, these terms are useful to classify snore sounds based on their excitation locations.

Generation of snoring sound is very similar to speech production mechanism [10]. Snore sound is produced by vibration of soft tissues in the upper airway, which acts as an acoustic filter during snoring sound production [1]. As for the speech signal, it is generated by vibrations of the vocal cords, which is filtered and transmitted through the upper airway and oral cavity [5]. Considering the similarity in production mechanism of speech and snoring sound, it is highly likely that there is acoustic similarity between the two sounds. Hence, we use RT-07 dataset [11] which includes speech, breath, noise, etc.,. to train UBM.

3. Correlation Feature Selection (CFS) and Random Forest Classifier

For the 2^{nd} sub-system, we employ Correlation Feature Selection (CFS) [12] to select the optimal subset from the 6373-dimension Open Smile feature data provided by organizers. The idea of CFS is to select the subset containing feature components which are highly correlated to the class labels and at the same time uncorrelated with each other. The non-informative feature components are eliminated due to their low correlation with class labels and the redundant feature components are also filtered out as they are highly correlated with one or some other components in the selected subset. The evaluation measurement is set as below:

$$M_S = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \tag{4}$$

Where S denotes a subset of k -feature components, r_{cf}: averaged feature-class cross correlation; r_{ff}: averaged feature-feature cross correlation. The algorithm is implemented with heuristic search mechanism. Best first strategy was found most effective in our experiments. It starts with single feature ranking to select the best ranked component. Then, we iteratively expand the size of subset until no increment was seen in the evaluation measurements. Particularly, the search will be terminated if five consecutive attempts to expand the subset are not successful.

4. Spectrogram and Convolutional Neural Network (CNN) Classifier

Convolutional neural networks (CNN) [13], which is a biologically inspired class of deep learning models provide the means

$$\begin{aligned}\Psi^{Bhat}\left(p_a \| p_b\right) \approx{}& \frac{1}{8}\sum_{i=1}^{M}\left\{\left[\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^a-m_i^u\right)\right]^T\left[\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{-1/2}\left(m_i^b-m_i^u\right)\right]\right\}\\ &+\frac{1}{2}\sum_{i=1}^{M}\mathrm{tr}\left[\left(\frac{\Sigma_i^a+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^a\right)^{-1/2}\left(\frac{\Sigma_i^b+\Sigma_i^u}{2}\right)^{1/2}\left(\Sigma_i^b\right)^{-1/2}\right]\\ &+\frac{1}{2}\sum_{i=1}^{M}\left\{\frac{\omega_i^u}{\omega_i^a}\frac{\omega_i^u}{\omega_i^b}\right\}-\sum_{i=1}^{M}\ln\left\{\omega_i^u\right\}-M \qquad (2)\end{aligned}$$

to extract local features from the spectrogram itself [14]. CNN uses relatively small-sized filters over a spectrogram patch and perform convolution. As convolution is done with adjacent bins in time and frequency, CNN has the capability to learn local feature maps [14]. Local features highlight continuity in time, continuity in frequency, or other more fluctuating local patterns. Hence, CNN is able to unfold a single spectrogram into many local feature maps and perform classification. The ability of learning local features allow CNN to capture acoustic characteristics independent of position in time or frequency. This explains that CNN is not learning event-dependent features, but rather useful local filters that reveal more independent aspects of sounds. We describe the process to compute spectrogram features and CNN structure in the following sections.

4.1. Generation of Spectrogram

The snore sounds are characterized by their sound intensity by a set of spectral parameters [15]. And, we generate a spectrogram for each snore sound clip. A spectrogram is a visual representation of the frequency spectrum over time, and the spectrograms of most sounds have several distinguishing features. In addition, the spectrogram contains frequency and amplitude information over time, and it is important acoustic information to distinguish different types of sounds. Before we compute spectrogram, we upsample the data set as some classes in the dataset has very few samples. We add the small random noise to the sample of small class to obtain more samples. Then, we compute the log power spectrogram using 40 ms frames with a 20 ms shift. The input into the neural networks was normalized to remove any variance.

4.2. Structure of Convolutional Neural Network (CNN)

In recent years, deep learning has not only permeated the image classification and speech recognition applications but also a powerful class of models for several classification and detection problems such as acoustic event detection (AED) [14], object detection [16], scene labeling [17] and house number digit classification [18]. Encouraged by these results, we employ CNN for snore sound classification. Sounds have global spectral patterns, as well as local properties such as being more transient or smoother in the time-frequency domain [14]. These can be exposed by using a model that exploits locality leading us to explore two different feature extraction strategies in the context of Convolutional neural networks (CNN) based deep learning. Hence, we employ CNN, a state of the art 2D feature extraction model, to exploit local structures, with log power spectrogram as input in our 3^{rd} sub-system for snore sound classification.

The architecture of our CNN model is as follows. The first Convolution layer takes the spectrogram as input. The model uses 2 Convolutional layers with Relu or Rectilinear units as activation, followed by Max Pooling the output. The resulting output is flattened before being fed to a fully connected neural layer which feeds the output to the final neural layer of classification which has the output neurons as the numbers of classes, using 'softmax' as activator.

Local acoustic characteristics of snore sounds in time-frequency spectrogram is learned in convolution layers. Each output neuron of convolution layer is connected to local regions in the input and each compute a dot product between inputs in the region and their respective weights [19]. Then, max-pooling operation identifies the important features in the region for each filter. This operation reduces the amount of parameters and computation in the network, and hence it controls overfitting. Finally, CNN stack features of all filters to produce the output volume. Hence, each output neuron of CNN corresponds to a small region in the input and CNN learns local characteristics of snore sound.

5. Database and baseline features

We use MUNICH-PASSAU SNORE SOUND CORPUS (MPSSC) which is provided by ComParE2017 challenge for snore sound classification based on their excitation location. Four classes are defined based on the VOTE scheme. V refers to Velum (palate), including soft palate, uvula, lateral velopharyngeal walls. O is for Oropharyngeal lateral walls which include palatine tonsils. T is for Tongue, including tongue base and airway posterior to the tongue base. And, E is for Epiglottis. The database contains audio samples of 843 snore events from 224 subjects. Please refer to [20] for details of the database. ComParE2017 provides the baseline feature set which contains 6373 static features [21] resulting from the computation of various functionals over low-level descriptor (LLD) contours.

6. Experiments

We conduct several experiments to investigate the effectiveness of the individual sub-systems and fused system. Firstly, we examine effect of integrating Bhattacharyya based GMM supervector to the baseline feature set. We use Support Vector Machine(SVM) for classification with the integrated feature set. Secondly, we investigate feature selection method, CFS, to improve the performance of the baseline system by using a subset of baseline feature set which is important for classification of 4 different snore sounds. Thirdly, we examine the deep learning method based on Convolutional Neural Network (CNN) to classify snore sounds. We use log power spectrogram as input features for CNN classifier. Finally, we fuse above three systems for snore sound classification. In all experiments, we use Unweighted Average Recall (UAR) [20] to measure the classification accuracies. The following are experiments with 3 subsystems as well as fused system.

6.1. Effect of using Bhattachrayya based GMM supervectors and SVM classifier

Each utterance is divided into 20ms frames with 10ms overlapping. Each frame is multiplied by a Hamming window to minimize signal discontinuities at the end of each frame. From each frame, we extract Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficient (LPCC) and Perceptual Linear Prediction Coefficients (PLPC) features. Each feature has 12 coefficients and their first derivatives. We form a feature vector for each frame by concatenating all three MFCC, LPCC and PLPC features. As each feature has a total of 24 coefficients, a feature vector of a frame has 72 coefficients. To formulate GMM-supervectors, we use maximum a posteriori (MAP) criterion to adapt a GMM model from a Universal Background Model(UBM) for each snore sound. We train Universal Background Model (UBM) via EM algorithm [22]. We use Rich Transcription 2007 [11] Evaluation dataset to train UBMs. This RT-07 dataset includes speech, room noise, paper flipping, door noies, breath sounds recorded during a meeting. We adapt the mean and covariance only. Once we have an adapted GMM model, we formulate Bhattacharyya based GMM-supervectors using the techniques mentioned in Section 2.2. Snore sound classification accuracies on development and test sets of Snore Sub-Challenge using Bhattacharyya based GMM supervectors (Bhat-GMM-Sup) together with 6373 Baseline Features (BLF) are presented in 3^{rd} row of Table 1. We use training set to train the snore sound models. The result in Table 1 shows that snore sound classification accuracy improves 7.7% absolute over the challenge baseline system for development set (3^{rd} row of Table 1) when we integrate Bhattacharyya based GMM supervectors (Bhat-GMM-Sup) to baseline features.

Table 1: *Unweighted average recall (UAR [%]) of individual sub-systems and fused system on development set and test set, (BLF = Baseline Features)*

Features	Classifier	Devel	Test
BLF [21]	SVM	40.6	58.5
BLF + Bhatt-GMM-Sup	SVM	48.3	52.4
BLF + CFS	RF	56.22	50.7
Spectrogram	CNN	45.4	49.3
Fused system		57.13	51.7

6.2. Correlation Feature Selection (CFS) and Random Forest classifier

In this experiment, we observe effect of feature selection on challenged provided 6373 Baseline Feature (BLF) set using Random Forest (RF) as classifier. Although the measurements such as MDL (Maximum Description Length), SU (Symmetric Uncertainty), or R (Relief) can be used as alternative evaluation measurement for feature selection, the simplest CFS shows most effective in our experiments. In our experiments, 25% of training data (excepting the class-3 with full set of 26 samples) was used in the CFS feature selection. The algorithm selects 53 feature components into the classification. The Random Forest is adopted as the classifier. A simple configuration of 9 trees depth forest, train 100 of them, use 5 random splits when training each weak learner is used in the evaluation. We achieve UAR of 56.22% and 50.7% for respectively for development and test sets as shown in 4^{th} row of Table 1.

6.3. Spectrogram and CNN classifier

In this section we conduct snore sound classification experiments using CNN classifier and log power spectrogram as input to CNN. Process to compute log power spectrogram is described in Section 4.1. CNN has 2 convolution layers, 2 MaxPooling layers and 1 softmax layer. Each convolutional layer is followed by a MaxPooling layer. The first convolution layer uses 32 initial convolution filters and a convolutional kernel of 5 rows and 5 columns. It is followed by 2 dimensional max pooling over 5x5 blocks. And, the second convolutional layer has 64 filters and kernel is 5*5. The 2nd convolution layer is followed by 2 dimensional max pooling over 2x2 blocks. Both Convolution layers uses Relu or Rectilinear units as activation. The last layer is a fully connected Dense neural layer with an output dimensionality of 4 as we are classifying 4 different types of snore sounds. The final Dense layer uses softmax as its activation function. The model uses the Categorical Cross Entropy as the loss function and the optimizer is first-order gradient-based optimization of stochastic objective functions (adam) with a learning rate of 0.0001. For our experiments we use 30 epochs and train our model using batch size of 250. Our CNN model is implemented using Keras [23] libraries in Python. With the above setup, we achieve the classification accuracy of 45.4% and 49.3% on development and test sets respectively as shown in 5^{th} row of Table 1. The result demonstrates that exploring local acoustic properties in time-frequency spectrogram is useful for snore sound classification.

6.4. System fusion

Finally, we fuse the 3 sub-systems to obtain the final snore sound classification results. The results in Table 1 show that CFS feature selection process is effective and this system gives the best performance among all individual sub-systems for development set. We employ majority voting using output scores of individual sub-systems. In case if voting is equal, we take the decision of the best system, CFS with RF, as the final classification decision. The fusion process gives the final classification UAR of 57.13% and 51.7% respectively for development and test sets as shown in 6^{th} row of Table 1. The fused system achieves 16.53% absolute improvement in terms of UAR over the challenge baseline system for development set.

7. Conclusion

In this paper, we have presented the 3 sub-systems for snore sound classification. In the first system, we examine acoustic distance measure of snore sounds and integrate the information of distance measure information to baseline features. In the second system, we investigate to select subset of features which is important to classify snore sounds from baseline feature set. In the third system, we explore local properties of sound characteristics for snore sound classification. The individual sub-systems show performance improvements over baseline system for development set. Fusion of the three sub-systems further improves the classification of 4 snore sound types. In future, we will explore multi-task deep learning approach that has benefit of more data for training by leveraging data from many tasks. We expect that the system will perform better for unseen test data.

8. References

[1] D. N. Fairbanks and S. Fujita, *Snoring and obstructive sleep apnea*. Lippincott Williams & Wilkins, 1994.

[2] K. Wilson, R. A. Stoohs, T. F. Mulrooney, L. J. Johnson, C. Guilleminault, and Z. Huang, "The snoring spectrum: acoustic assessment of snoring sound intensity in 1,139 individuals undergoing polysomnography," *CHEST Journal*, vol. 115, no. 3, pp. 762–770, 1999.

[3] A. Yadollahi and Z. Moussavi, "Automatic breath and snore sounds classification from tracheal and ambient sounds recordings," *Medical engineering & physics*, vol. 32, no. 9, pp. 985–990, 2010.

[4] R. Nonaka, T. Emoto, U. R. Abeyratne, O. Jinnouchi, I. Kawata, H. Ohnishi, M. Akutagawa, S. Konaka, and Y. Kinouchi, "Automatic snore sound extraction from sleep sound recordings via auditory image modeling," *Biomedical Signal Processing and Control*, vol. 27, pp. 7–14, 2016.

[5] M. Shokrollahi, S. Saha, P. Hadi, F. Rudzicz, and A. Yadollahi, "Snoring sound classification from respiratory signal," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 3215–3218.

[6] W. Duckitt, S. Tuomi, and T. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiological measurement*, vol. 27, no. 10, p. 1047, 2006.

[7] K. Qian, Z. Xu, H. Xu, Y. Wu, and Z. Zhao, "Automatic detection, segmentation and classification of snore related signals from overnight audio recording," *IET Signal Processing*, vol. 9, no. 1, pp. 21–29, 2015.

[8] T. Mikami, Y. Kojima, M. Yamamoto, and M. Furukawa, "Automatic classification of oral/nasal snoring sounds based on the acoustic properties," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 609–612.

[9] C. H. You, K. A. Lee, and H. Li, "Gmm-svm kernel with a bhattacharyya-based distance for speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1300–1312, 2010.

[10] A. M. Alencar, D. G. V. da Silva, C. B. Oliveira, A. P. Vieira, H. T. Moriya, and G. Lorenzi-Filho, "Dynamics of snoring sounds and its connection with obstructive sleep apnea," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 1, pp. 271–277, 2013.

[11] N. Spring, "Rich transcription meeting recognition," 2007.

[12] M. A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand: University of Waikato, 1998.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," vol. 86, no. 11. IEEE, 1998, pp. 2278–2324.

[14] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.

[15] J. Sola-Soler, R. Jane, J. Fiz, and J. Morera, "Variability of snore parameters in time and frequency domains in snoring subjects with and without obstructive sleep apnea," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*. IEEE, 2006, pp. 2583–2586.

[16] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[17] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," vol. 35, no. 8. IEEE, 2013, pp. 1915–1929.

[18] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3288–3291.

[19] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap dataset." in *Twenty-Ninth IAAI Conference*, 2017.

[20] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring."

[21] F. Eyben, "Real-time speech and music classification by large audio feature space extraction," 2015.

[22] L. N. R. D. Dempster, A.P., "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[23] "Keras: Deep learning library for tensorflow and theano," accessed: 2016-09-30.