Separation Of Speech From Noise Challenge

NagaChaitanya Vellanki
vellanki@stanford.edu

December 14, 2012

1 Introduction

The goal of this project is to implement the methods submitted for the PASCAL CHiME Speech Separation and Recognition Challenge[1] [1]. In particular, estimating the spectrographic mask using SVM for missing feature methods [15] of noise compensation.

2 CHiME

The main task in the CHiME challenge is to recognise the letter and digit in each noisy utterance.The dataset consists of utterances of simple sentences by 34 speakers (18 male and 16 female) in a domestic environment in the presence of noise sources of a typical family home: two adults and two children, TV, footsteps, electronic gadgets(laptops and game console), toys, some traffic noise from outside and noises arriving from a kitchen via connecting hallway. The recordings were made using a mannequin with built-in left and right ear simulators that record signals that are an approximation of the acoustic signals that would be received by the ears of an average adult listener. The sentences consist of simple six word commands of the following form:

Command format:
($command = $color $preposition $letter $number $adverb)
where each word can have the following alternatives,
$command = bin | lay | place|set;
$colour = blue | green | red | white;
$prep = at | by | in | with;
$letter = A | B | C | ... | U | V | X | Y | Z;
$number = zero | one | two ... seven | eight | nine;
$adverb = again | now | please | soon;

Example commands:
lay blue by H five again
lay blue in T four again

The training data consists of 3600 stereo 16 bit WAV files (600 utterances at 6 different SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB)) at 16 kHz or 48 kHz. Each WAV file contains a single noisy utterance. The noise background can have multiple sources but not more than 4 active sources at a time. The speech and noise backgrounds are two channel signals. The SNR defined as

$$SNR_{dB} = 10\log_{10}\left(\frac{E_{s,l}+E_{s,r}}{E_{n,l}+E_{n,r}}\right)$$

where l, r refer to the left and right channels and s,n are speech and noise backgrounds. E is the energy which is the sum of the squared sample amplitudes measured for the speech or background signals between the start and end points of the utterance.

The data set also has 17,000 files containing 500 utterances of each of the 34 speakers to train acoustic speech models. These utterances were provided with reverberation but free of additive noise. Additional 6 hours of background noise data for train background models. The test set is similar to the training set (600 utterances at 6 different SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB)) at 16 kHz. There is no overlap between the backgrounds of the test set and the noisy background data. Under the challenge guidlines, **models should not take advantage of the SNR labels and should not exploit the fact that same utterances are used at different SNR**.

3 Representing data using spectrograms

The methods used in this project operate on log-mel spectrograms[2] of the utterances. These log-mel features are computed from WAV files using HCopy of the HTK toolkit with TARGETKIND is set to FBANK_0. The log-mel spectrograms of Speaker 34 for Command: lay blue in T four again are shown in Figure 1, 2 and 3

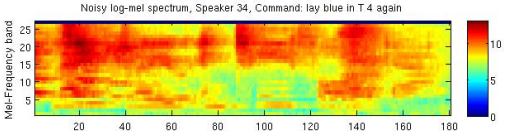

Figure 1: command with a child's voice in background at 0 dB SNR

[1] http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html

[2] A spectrogram is a two-dimensional representation of a speech signal. In spectrogram time is displayed on x-axis and the frequency on y-axis. Each time-frequency location in the spectrogram represents the power of the signal. In log-mel spectrogram, time is displayed on x-axis and logarithm of the output of k^{th} mel filter on y-axis. See section 2.2 of [12] for more details on spectrogram and variants

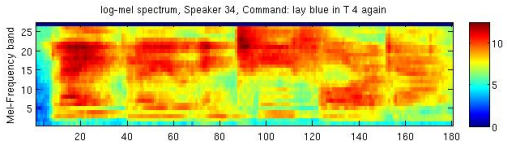

Figure 2: command with no background noise

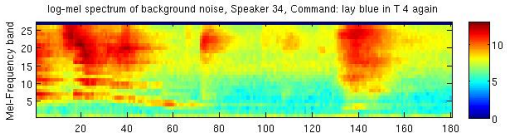

Figure 3: background noise at 0 dB SNR

4 Spectrographic Mask Estimation

Spectrographic mask estimation methods divide the observed log-mel spectral features into speech, noise dominated regions. The speech dominated time-frequency components of are considered reliable estimates of clean speech. *S(t,f)* is the clean speech that could have been observed if the signal was not corrupted with noise. The noise dominated time-frequency components *N(t,f)* are considered unreliable, they only provide a upper bound on the speech values [2] $N(t,f) \geq S(t,f)$ .We can see that clean speech information is missing in unreliable components. The spectrographic masks are used in Missing Feature methods of noise compensation for speech recognition in order to identify unreliable components. Missing feature methods were shown to be very successful at compensating noise when the spectrographic mask labeling every time-frequency location as reliable or unreliable is known [15][16].In missing feature methods the recognition is then performed using the reliable components or by reconstructing the unreliable components prior to the recognition.

4.1 Oracle Mask

The 'oracle mask' [5] can be constructed by comparing the log-mel spectral features of the clean speech S with the added noise N. The reliability of time-frequency cell is given by [3]

$$M(k,j) = \begin{cases} 1 \stackrel{def}{=} reliable & S(k,j) \geq N(k,j) - \theta \\ 0 \stackrel{def}{=} unreliable \end{cases}$$

where k is the frequency band, j is the time-frame and $\theta = -2$ dB is the fixed mask threshold.

The oracle masks were computed for all utterances across

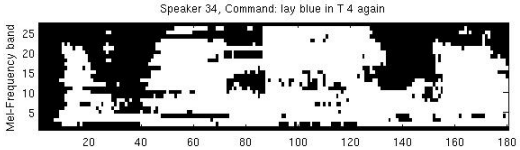

Figure 4: oracle mask with a threshold of -2 dB SNR, black regions in the mask denote unreliable features

SNRS using the clean speech, background noise files provided in training set. These oracle masks will be used to provide reliability labels for the features of the SVM classifier.

4.2 Feature for the SVM

'Subband energy to subband noise floor ratio', 'Subband energy to fullband noise floor ratio', 'Flatness', 'Subband energy to full band energy ratio', 'Kurtosis', 'Spectral-subtraction-based SNR estimate' are used as the features for the classifier. Missing eature methods do not make any assumptions about the nature of the corrupting noise so the mask estimation process should also be be free of assumptions about the noise. The above features make minimal assumptions about the background noise and rely only on the characteristics of the speech signal. The details of the features will be described here briefly (refer to [7][12] more details):

4.2.1 Subband energy to full band energy ratio

Subband energy to full band energy ratio is the log ratio of the energy in subband to the overall frame energy. As background noise is added to speech, the spectral shape changes as a function of the spectral characteristics of the noise. Subband energy to full band energy ratio is a measure of the effect of background noise on a particular subband and on the overall frame.

4.2.2 Subband energy to subband noise floor ratio

Noise floor of a the noise-corrupted speech signal is useful for estimating the SNR. The energies of all frames of an subband are put into a histogram and the lower peak is found. The energy bin in the histogram corresponding to this peak value is considered as noise floor. The ratio of the energy of a subband of a frame to the noise floor in the subband will help determine that a specific spectrographic location has been corrupted by noise.

4.2.3 Subband energy to fullband noise floor ratio

The energies of all frames of an utterance are put in a histogram and the lower energy peak is found. The energy bin in the histogram corresponding to this peak value is the noise floor of the noisy speech signal. The ratio of the energy of a subband of a frame to the noise floor of the noisy speech signal will help determine that a specific spectrographic location has been corrupted by noise.

4.2.4 Spectral-subtraction-based SNR estimate

The SNR estimate used to compute the oracle masks. Including SNR estimation was shown to provide improvement over baseline recognition in [13].

4.2.5 Flatness

Flatness is the variance of subband energy in a neighborhood of spectrographic locations around a given pixel. Noise-corrupted spectrographic locations have a lower variance than cleaner ones. Flatness is given by the following equation

$$\sigma^2_{flat}(n,\omega_i) = \frac{1}{9} \sum_{k=i-1}^{i+1} \sum_{j=n-1}^{n+1} (s(j,\omega_k) - \mu_s(n,\omega_i))^2$$

for a 3×3 neighborhood of pixels where $s(n, \omega_i)$ represents the subband energy of frame n and subband ω_i, and $\mu_s(n, \omega_i)$ is the mean of the subband energy values in 3×3 neighborhood around frame n and ω_i

4.2.6 Kurtosis

Kurtosis is defined as

$$K_x = \frac{E\{x^4\}}{\{E\{x^2\}\}^2}$$

where the expectations are calculated for each subband.

4.3 SVM Mask Estimation

An SVM classifier is trained for each of the F(26) mel-frequency bands for each of the 34 speakers using LIBSVM [8] on 5400 frames randomly extracted from the utterances of the particular speaker in the training set across different SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB)), with a total of 26×34 models. Reliability labels used in training were derived from the oracle mask of the utterances obtained from the clean speech and background noise data. Each classifier used the same set of single-frame based features 'Subband energy to subband noise floor ratio', 'Subband energy to full-band noise floor ratio', 'Flatness', 'Subband energy to full band energy ratio', 'Kurtosis', 'Spectral-subtraction-based SNR estimate' features derived from the noisy mel-features along with the noise mel-features. The features were normalized to mean 0 and variance 1 before training the SVM. The SVM was trained using the RBF Kernel and the hyperparameters c, γ were chosen using grid search in $A \times A$ where $A = \{2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}\}$ by doing a 5-fold cross validation on additional held-out 600 frames. This setup was used in [2] for SVM mask estimation. Each model was tested on 5000 additional held-out frames in the training set. The results for each of the 26×34 model were captured, only results for the speaker 33, 34 on some randomly selected utterances will be described in this report and the results for rest of the speakers will be handed in along with the report. The SVM estimated masks were obtained by testing the above trained models on utterances at SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB).**Figure 5-10, SVM Estimated Masks at different SNR along with the oracle mask of threshold at -2 dB SNR for Speaker 33, Command: lay blue by H 5 again**

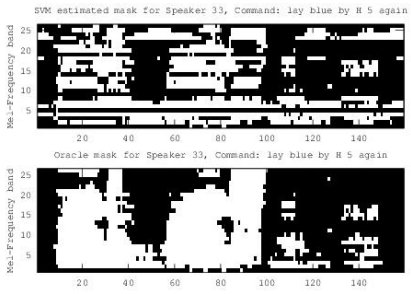

Figure 5: -6 dB SNR

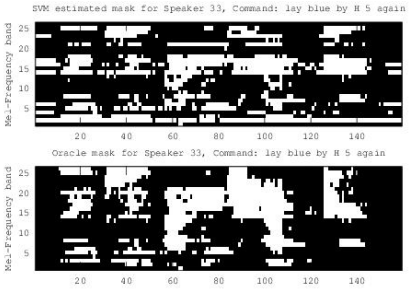

Figure 6: -3 dB SNR

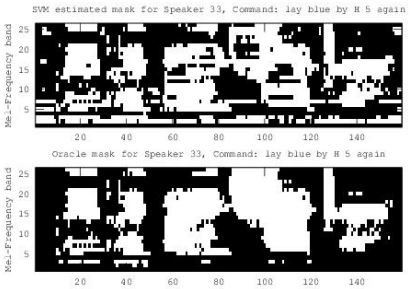

Figure 7: 0 dB SNR

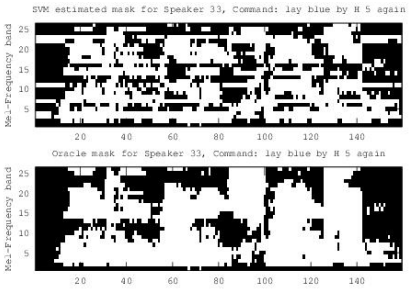

Figure 8: 3 dB SNR

Figure 9: 6 dB SNR

5 Evaluation and Experiments

The performance of the mask estimated by the SVM classifier can be evaluated in two ways

1. The classification accuracy of the estimated mask compared to the oracle mask.
2. The improvement in recognition accuracy achieved by

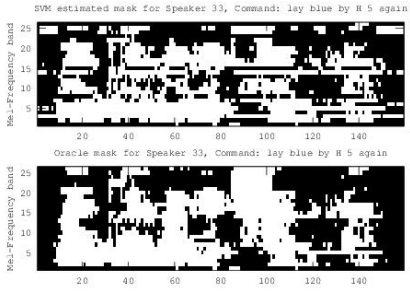

Figure 10: 9 dB SNR

using the classifier-generated masks in missing feature methods.

In this project,the performance of the mask estimated by the classifier is evaluated by comparing it to the oracle mask as described in [12].

5.1 Evaluation

There are two types of errors the classifier can make 'miss' and 'false alarm'. A 'miss' can be defined as incorrect labeling of unreliable spectrographic location as reliable and 'false alarm' as incorrect labeling of a reliable spectrographic location as unreliable. Similarly, there are two types of correct identifications the classifier can make: 'hit' and 'correct rejection'.A 'hit' can be defined as correct labeling of a unreliable spectrographic location and 'correct rejection' as correct labeling of a reliable spectrographic location. The classifier is considered optimal if it maximizes hits and minimizes false alarms. As seen in Figure 11, the classifier clearly needs more information to correctly identify reliable spectrographic locations as SNR information cannot be used in the models. Further experimentation can be done by adding additional features like Harmonic [2], aperiodic part of the harmonic decomposition [6] , long term energy estimate [2], gain factor [2], VAD [14], Comb filter ratio [7][12], Autocorrelation peak ratio [7][12] to the classifier and also by including neighboring $N \times N$ features around a spectrographic location as there is some correlation between a reliable spectrographic location and its neighbors[12].

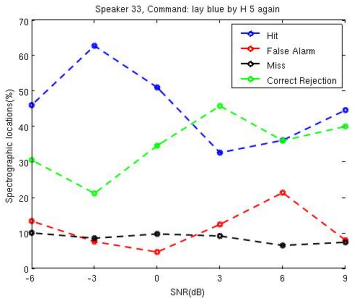

Figure 11: Percentage Hit, Miss, False Alarm, Correct Rejection for Speaker 33, Command: lay blue by H 5 again at SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB)

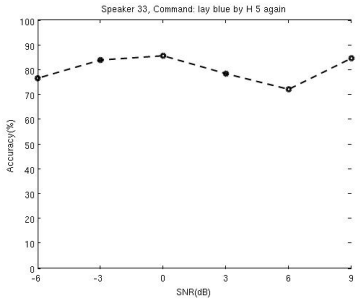

Figure 12: Classification Accuracy for Speaker 33, Command: lay blue by H 5 again at SNR (-6 dB, -3 dB, 0 dB, 3 dB, 6 dB, 9 dB)

5.2 Experiments

5.2.1 Varying Training set size

The classifier was trained training set with varying training set sizes from 5400 to 11400 in steps of 1000 for Speaker 34, Command: lay blue in T 4 again at 0 dB SNR and the masks were obtained for each training set size. There was little improvement in the accuracy but the original problem of correctly identifying reliable and unreliable spectrographic locations remained.

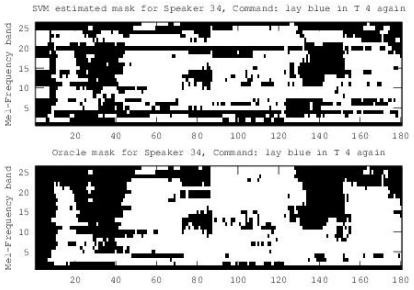

Figure 13: Mask obtained after training with 11400 frames and tested with Speaker 34, Command: lay blue in T 4 again at 0 dB SNR

5.2.2 Spectral-based-subtraction SNR estimate as feature

The classifier was trained without, with the Spectral-subtraction-based SNR estimate feature to see the improvement in classification accuracy. The classification accuracy improved when SNR estimate was one of the feature as stated in [13]. This experiment also shows that the classification accuracy is not exclusively controlled by the SNR estimate as shown in Table 1.

Table 1: **classification accuracy of SVM for speaker 34, across 26 mel-frequency bands**

	Without SNR estimate		With SNR estimate	
	Train	Test	Train	Test
1	100	97.84	99.79	99.6
2	75.46	73.74	99.77	99.42
3	70.24	70.12	99.75	99.38
4	65.96	65.8	99.77	99.04
5	63.07	62.2	99.87	99.56
6	71.64	65.92	99.79	99.34
7	69.37	68.72	99.85	99.74
8	68.33	66.44	99.74	99.4
9	61.90	62.46	99.74	99.48
10	70.31	65.52	99.77	99.26
11	77.57	66.34	99.70	99.3
12	66.40	64.78	99.77	99.08
13	71.66	69.28	99.90	99.48
14	70.88	68.16	99.64	98.88
15	69.48	67	99.75	99.64
16	69.33	67.68	99.72	98.38
17	70.22	69.56	99.44	98.9
18	66.05	66.74	99.68	99.2
19	68.14	64.46	99.83	99.54
20	65.96	62.36	99.59	99.56
21	70.53	68.86	99.81	99.14
22	76.62	75.82	99.61	98.96
23	82.42	80.64	99.81	98.4
24	83.64	83.28	99.68	99.56
25	83.16	82.14	99.81	99.44
26	79.59	78.6	99.92	99.28

6 Future work

1. Converting features from log-spectra to cepstral domain. Since log-spectra and cepstra are related by a linear transform, a solution for converting from log-spectra to cepstral domain has been described in [10].
2. Add additional features like Harmonic [2], aperiodic part of the harmonic decomposition [6] , long term energy estimate [2], gain factor [2], VAD [14], Comb filter ratio [7][12], Autocorrelation peak ratio [7][12] in the classifier
3. Use the spectrographic mask obtained using the SVM classifier in missing feature compensation methods of speech recognition and run the baseline recognizer system to compare the results with submissions

7 Acknowledgements

I would like to thank **Andrew Maas**[3], Stanford University for this project suggestion and for helping through the project, **Jort Florent Gemmeke**[4], ESAT-PSI speech group, KU Leuven, Belgium.for providing MDT Tools package to understand the mask estimation process. Special thanks to **Mike Seltzer**[5], Speech Technology group, Microsoft Research for explaining about extracting training data for the SVM classifier using the log-mel features and reliablity labels from the oracle mask. His thesis [12] has been very useful in understanding the details of the mask estimation process.

[3] http://ai.stanford.edu/~amaas/

[4] http://www.amadana.nl/

[5] http://research.microsoft.com/en-us/people/mseltzer/

8 References

[1] The pascal chime speech separation and recognition challenge (2011) by J Barker, H Christensen, N Ma, P Green, E Vincent.

[2] Kallasjoki, H., Keronen, S., Brown, G. J., Gemmeke, J. F., Remes, U., Palomaki, K. J., 2011. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 5863.

[3] J. Gemmeke, B. Cranen, and L. ten Bosch, On the relation between statistical properties of spectrographic masks and recognition accuracy, in SPPRA- 2008, 2008, pp. 200206.

[4] Jort F Gemmeke, B Cranen (2009) TR02 : State dependent oracle masks for improved dynamical features http://arxiv.org/abs/0903.3198

[5] Christophe Cerisara, Sebastien Demange, and Jean-Paul Haton, On noise masking for automatic missing data speech recognition: A survey and discussion,Comput. Speech Lang., vol. 21, no. 3, pp. 443457,2007.

[6] H. Van hamme, Robust speech recognition using cepstral domain missing data techniques and noisy masks, in Proc. ICASSP, Montreal,Quebec, Canada, May 2004, pp. 213216.

[7] M. Seltzer, B. Raj, and R. Stern, A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition, Speech Communication, vol. 43, no. 4, pp. 379393, 2004.

[8] C. Chang and C. Lin, LIBSVM: a library for support vector machines, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9020

[9] Van Hamme, H.; , "Handling Time-Derivative Features in a Missing Data Framework for Robust Automatic Speech Recognition," Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on , vol.1, no., pp.I, 14-19 May 2006.

[10] Van hamme, H., Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain, Proc. Eurospeech, Geneva, Sept. 2003, pp. 3089-3092.

[11] VOICEBOX: Speech Processing Toolbox for MATLAB, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[12] M. L. Seltzer, Automatic Detection of Corrupted Speech Features for Robust Speech Recognition, Master's Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, May, 2000

[13] Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integratedstudy. Proc.Eurospeech'99.

[14] J. Ramrez, J. Gorriz, J. Segura, C. Puntonet, and A. Rubio, Speech/non-speech discrimination based on contextual information integrated bispectrum LRT, in IEEE Signal Processing Letters, vol. 13, no. 8, 2006, pp. 497500.

[15] M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data, Speech Commun., vol. 34, pp. 267285, 2001

[16] Raj, B., Reconstruction of Incomplete Spectrograms for Robust Speech Recognition, Ph.D.Dissertation, Carnegie Mellon University, May 2000.