

EVALUATION OF NOISE ROBUST FEATURES ON THE AURORA DATABASES

Xiaodong Cui, Markus Iseli, Qifeng Zhu, and Abeer Alwan

Department of Electrical Engineering
University of California, Los Angeles, CA 90095
Email: {xdcui, iseli, qifeng, alwan}@icsl.ucla.edu

ABSTRACT

In this paper, we evaluate our noise robust feature extraction algorithms on the Aurora 2 and the German part of Aurora 3. Several algorithms are introduced and evaluated to deal with the noisy speech signals including our previous noise robust techniques used with Aurora 2, and new approaches evaluated with Aurora 3. Since there exist some differences between the two databases, modifications of front-end modules are needed. For Aurora 2, the average error rate reduction is 47% for clean training and 12% for multicondition training compared with the new baseline with endpoint detection. In Aurora 3, we obtain 17%, 27% and 53% error rate reduction for the well-matched, medium-mismatched and high-mismatched cases, respectively.

1. INTRODUCTION

Aurora 2 [1] and Aurora 3 [2] are two databases used to test different front-end feature extraction algorithms in noisy environments for distributed speech recognition (DSR). For DSR, it is crucial to explore noise robust features that maintain high performance at low SNR or at least drop gracefully as the SNR decreases.

In this paper, we introduce our previous work on Aurora 2 [3] and new progress on Aurora 3. Although both databases are digit strings recorded in noisy environments, there are some differences between them. In Aurora 2, data are divided into clean training and multicondition training sets while the testing sets contain only those files from the multicondition case. Various types of noise are manually added to the clean speech at different SNR levels. In Aurora 3, data are categorized into well-matched, medium-mismatched and high-mismatched cases, and the signals are recorded using two types of microphones. For each case, different noise levels are included in both training and testing sets. Therefore, for Aurora 2, the noise level difference is the main factor for data mismatch while for Aurora 3, channel mismatch contributes more. For the reason mentioned above, we modify the front-end modules applied to the two databases in order to get satisfactory performance.

For both databases, 18-state word-based HMM models (including two dummy states) are adopted. We use the HTK platform provided by the databases for training and recognition while we implement our algorithms in the front-end modules. The final 39-dimension feature vectors include 12 cepstral coefficients without C_0 plus log energy and their first, second order derivatives.

The rest of this paper is organized as follows: we describe the algorithms applied in both databases in Section 2. In Sections 3 and 4, front-end realizations and experimental results on Aurora 2 and Aurora 3 are described, respectively. Finally, we present conclusions in Section 5.

2. ALGORITHMS IN NOISE ROBUST FEATURE EXTRACTION

In this section, we introduce the algorithms applied in noise robust feature extraction on Aurora 2 and Aurora 3.

2.1. Speech-nonspeech detection

Although there are silence and short pause models in the HMM configuration to deal with nonspeech signals, it is still necessary and important to detect speech signals from nonspeech. This is because in DSR, the transmission bit rate during the nonspeech period can be reduced and more crucially, it has been shown beneficial for speech recognition, especially in noise, if only the detected speech signal is considered in the recognition process. Hence, a robust and effective speech-nonspeech detector is highly needed in this noise-prone task.

In Aurora 2 clean training condition, we detect speech from nonspeech by track the pitch for voiced speech. In Aurora 3 and Aurora 2 multicondition training case, we design a speech-nonspeech detector as shown in Fig. 1. Normally, when the SNR is tolerable, energy is a good choice as a parameter for speech-nonspeech detection [4]. But as the SNR decreases, energy becomes vulnerable. In the Aurora databases investigated in this paper, the SNR ranges from about -5dB to 30dB. Such a wide range of SNRs can not be handled by energy only. Spectral entropy is robust to identify voiced sound in case of low SNR which acts as a complement to energy [5]. In Fig. 1, input signals are sent into a wait-word-pause three-state model with estimated online SNR using log energy. Under certain constraints, the model outputs the status of the current frame. In the mean time, artificially generated white noise is added to the input signal in a controlled manner. Spectral entropy is computed afterwards which is combined together with the three-state model output to give the speech or nonspeech tag for the current frame. In the post-processing part, we make a final decision by discarding those speech or nonspeech intervals with too few frames and concatenating those satisfying certain criteria.

2.2. Variable frame rate analysis (VFR)

From the speech perception point of view, the transition from consonants to vowels contains important information. Variable frame rate (VFR) analysis [6] attempts to capture more detailed dynamic spectral characteristics in the transition part of the speech by assigning more frames, with shorter frame shift, and sparse frames within the steady-state segments. The selection of frame rates is based on the derivatives of weighted log energy Euclidean MFCC distances. In the transition portion of the speech, Euclidean MFCC

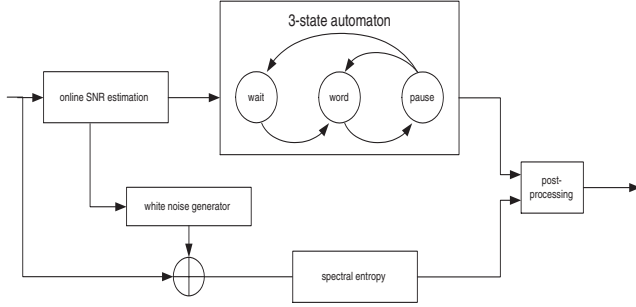


Fig. 1. Speech-nonspeech detector

distances show faster changes than the steady-state. In this paper, we choose the frame shift as a multiple of 2.5 ms and on average, the frame rate is less than 100 frames per second.

2.3. Peak isolation (PKISO)

In the presence of noise, peaks of the speech spectrum are critical for sound perception while valleys are easily corrupted by noise. Thus, it makes sense to choose the spectral peaks as robust features. The peak isolation approach introduced in [7] enhances peaks by liftering in the cepstral domain and IDCT back to the spectral domain. It is then followed by half wave rectification.

2.4. Peak-to-valley ratio locking

The peak-to-valley ratio locking method is motivated from the fact that even in a very noisy condition (e.g. SNR = 0dB), the formants almost remain unchanged compared to the valleys which are buried by the noise spectrum, and their magnitude levels are almost determined by noise spectral magnitude. Hence, there arises a mismatch between features extracted from clean and noisy speech. Clean speech has a higher peak-to-valley ratio than noisy speech. Our peak-to-valley ratio proposed in [3] "locks" the ratio by setting the highest peak to a fixed value (10 for this paper) and scaling all the other Mel-filter outputs.

2.5. Harmonic demodulation (HD)

Harmonic demodulation [8] aims at reducing the difference between clean and noisy speech spectrum especially in the inter-harmonic valleys.

The LTI speech production model is viewed as amplitude modulation in the frequency domain, with the excitation spectrum being the carrier and the spectrum of the vocal tract transfer function being the modulator. Non-coherent demodulation with nonlinear envelope detection is used to recover the spectrum of the vocal tract transfer function where only the maximum value is calculated as the output of demodulation process[4]. Compared with linear demodulation, this nonlinear envelope detection algorithm shows robustness when the noise level is lower than the clean speech peaks. Envelopes of the speech spectra, instead of the speech spectra themselves, are used to compute the MFCCs.

2.6. Spectral subtraction

Spectral subtraction is a method aiming at removing the noise spectrum from the corrupted speech spectrum [9]. In this tech-

nique, additive noise is assumed when its spectrum is estimated. During the estimation, the noise signal magnitude $|N(e^{j\omega})|$ is approximated by the average of first M frames which are assumed to be background noise or in an adaptive way

$$|N_{new}(e^{j\omega})| = \alpha|N_{old}(e^{j\omega})| + (1 - \alpha)|X_i(e^{j\omega})|$$

where i is the current frame index and α is chosen to be around 0.9. The adaptation is applied only in the nonspeech part of the signals using the information provided by the speech-nonspeech detector.

After the estimation, the noise spectrum magnitude is subtracted from that of noisy speech

$$|\hat{S}(e^{j\omega})| = \begin{cases} |X(e^{j\omega})| - |N(e^{j\omega})| & \text{if } |X(e^{j\omega})| > |N(e^{j\omega})|, \\ \beta|N(e^{j\omega})| & \text{otherwise.} \end{cases}$$

where β is the floor value. In this paper, we set β to 0.33.

2.7. Cepstral domain RASTA filtering

For the highly mismatched case in Aurora 3, there exists obvious channel mismatch for different microphones of the training and testing data set. This channel mismatch effects can be alleviated using a following RASTA bandpass filter in the cepstral domain [10]:

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})}$$

2.8. Other back-end model considerations

In addition to the algorithms mentioned above, we also apply some parameter tuning in the back-end HMM model which includes increasing the silence model variance and setting the penalty factor in the recognition process to a nonzero value. For those silence and short pause models trained in clean data, the variance is much smaller than those obtained from multicondition training. For all cases, except the multicondition training case in Aurora 2, we multiply the silence model variance by a factor of 1.1~1.2. In addition, when speech-nonspeech detector is added, the deletion error increases. In this case, we set the insertion penalty factor p in the HTK recognition platform to a positive value of 10~15 to balance the insertion and deletion errors.

3. IMPLEMENTATION OF FRONT-END MODULES

As mentioned in Section 1, the clean training condition in Aurora 2 corresponds more to SNR mismatch in spite of the different channel characteristics in set C. For Aurora 3 and multicondition training condition in Aurora 2, similar SNR levels appear in both the training and testing sets. So channel mismatch becomes more dominant. We find that SNR and channel mismatches in training and testing sets in Aurora databases lead to different performances of the algorithms. Some of the algorithms contribute more in the SNR mismatch case while others are good at channel mismatch. So with respect to the two different mismatch conditions, we make minor modifications in the implementation of the front-end modules.

The front-end feature extraction module for Aurora 2 clean training condition (FEAT1) is described in Fig. 2. After the speech-nonspeech detection, we use VFR to chose the speech frames.

Harmonic demodulation is applied in the spectral domain whose outputs are sent into Mel-filter bank and their log magnitudes are computed. Peak isolation and peak-to-valley ratio locking are used right after the DCT. For the nonspeech frames, we choose one frame every 25 ms and just compute standard MFCCs. Finally, RASTA filters all the selected speech and nonspeech frames.

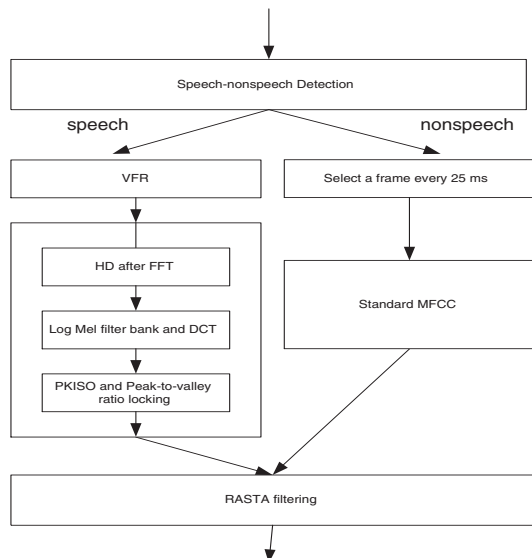


Fig. 2. Feature extraction module for Aurora 2 clean training (FEAT1).

In Aurora 3 and the multicondition training case of Aurora 2 (FEAT2), there is a slight difference from the front-end module introduced above. During the nonspeech segments, we estimate the noise spectrum from the first 15 frames or compute it adaptively in every nonspeech segment. The estimated noise spectrum is subtracted from the spectrum of speech segments which in turn are fed into harmonic demodulation after FFT. Only Peak-to-valley ratio locking is used after log Mel filter bank and DCT. As in Aurora 2 clean training, RASTA filtering is applied in the last step. Fig. 3 gives the detailed structure.

4. EXPERIMENTAL RESULTS

The performances of our proposed algorithms on noise robust feature extraction are given in Table 1 for Aurora 2 multicondition training, Table 2 for Aurora 2 clean training and Table 3 for Aurora 3 German. Since the new Aurora 2 baseline using endpoint detection [11] is different from that with the Aurora 2 CDs, error rate reductions in this paper are obtained by comparing our results with the new baseline. Note that the overall performance is calculated using the Excel form provided by D. Pearce [11] which assigns different weights to the performance of Sets A, B and C.

As shown in the table, the absolute overall accuracies for Aurora 2 clean training and multicondition training are 81.23% and 89.08%, respectively, which correspond to 47% and 12% error rate reductions. If we compare with the old Aurora 2 baseline without endpoint detection, the error rate deductions are 53% and 19%.

For the German part of Aurora 3, absolute accuracies are 92.69%, 86.24% and 87.28% for well-matched, medium-mismatched and

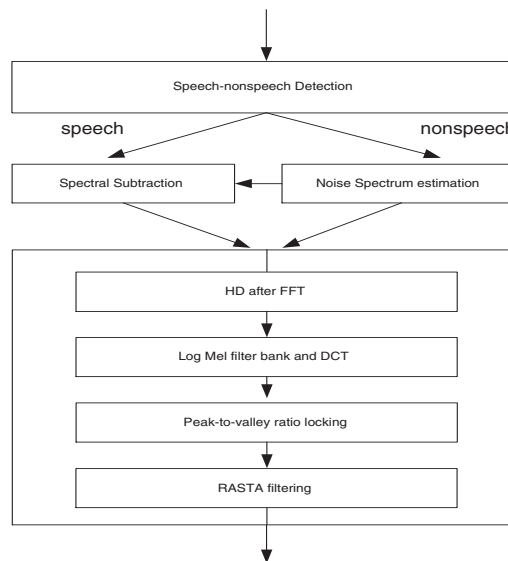


Fig. 3. Feature extraction module for Aurora 3 and Aurora 2 multicondition training (FEAT2).

high-mismatched cases. The error rate reductions are 17%, 27% and 53%, respectively.

	wm	mm	hm	0.4W+0.35M+0.25H
Absolute accuracy	92.69	86.24	87.28	89.08
Word error rate	7.31	13.76	12.72	10.92
Relative improvement	16.93	27.43	52.59	29.52

Table 3. Results on German of Aurora 3 in well-matched (wm), medium-mismatched (mm) and high-mismatched (hm) cases.

Experiments show that FEAT1 doesn't work as well in the multicondition training case of Aurora 2 and Aurora 3. This is mainly for the reasons discussed previously. It is interesting to note that in the multicondition training case, FEAT2 results in the highest improvement over FEAT1 in Set C where channel mismatch is the key issue. The average accuracy improves from 82.91% [3] to 87.25%.

5. CONCLUSIONS

In this paper, we evaluate our noise robust feature extraction algorithms on Aurora 2 and Aurora 3 (German only) databases. To handle SNR mismatch and channel mismatch issues, we apply modified versions of the same techniques. Compared with the new baseline, we obtain 47% and 12% error rate reduction on Aurora 2 clean training and multicondition training, respectively, and 30% on average for the Aurora 3 German database.

Aurora 2 Multicondition Training - Results(%)															
	A					B					C			Overall	Percentage Improvement
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average		
Clean	98.86	98.61	99.11	99.07	98.91	98.86	98.61	99.11	99.07	98.91	98.89	98.61	98.75	98.88	16.09
20dB	98.37	98.28	98.42	98.15	98.31	97.61	98.04	98.36	98.27	98.07	97.76	97.67	97.72	98.09	15.66
15dB	97.94	97.25	98.15	97.35	97.67	96.41	97.04	97.14	97.47	97.02	97.27	96.49	96.88	97.25	15.87
10dB	95.36	94.80	96.33	94.23	95.18	93.18	94.68	94.66	95.37	94.47	93.37	92.87	93.12	94.49	2.27
5dB	90.64	87.64	90.75	86.92	88.99	84.22	86.94	87.86	88.34	86.84	87.32	83.77	85.55	87.44	7.40
0dB	74.58	66.20	73.31	69.70	70.95	62.24	65.51	72.68	70.87	67.83	65.04	60.96	63.00	68.11	19.27
-5dB	39.18	31.32	30.69	35.59	34.20	29.72	32.20	36.81	33.88	33.15	32.90	29.71	31.31	33.20	11.02
Average	91.38	88.83	91.39	89.27	90.22	86.73	88.44	90.14	90.06	88.84	88.15	86.35	87.25	89.08	
Improv.	23.39	-0.38	20.28	11.00	13.57	0.51	6.98	16.80	26.81	12.78	15.67	-0.11	7.78		12.09

Table 1. Multicondition training results on Aurora 2 database using FEAT2

Aurora 2 Clean Training - Results(%)															
	A					B					C			Overall	Percentage Improvement
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average		
Clean	98.62	98.31	98.42	98.49	98.46	98.62	98.31	98.42	98.49	98.46	98.28	98.04	98.16	98.40	-101.32
20dB	96.68	97.25	97.20	96.76	96.97	97.18	96.95	97.32	97.50	97.24	94.87	95.74	95.31	96.75	29.25
15dB	94.50	96.22	95.65	94.60	95.24	95.64	95.71	95.85	95.77	95.74	91.19	93.14	92.17	94.83	53.88
10dB	89.62	92.90	91.38	87.87	90.44	91.25	91.14	91.02	90.65	91.02	82.04	85.19	83.62	89.31	60.66
5dB	76.33	83.01	81.93	75.10	79.09	75.47	79.81	80.55	77.23	78.27	59.66	69.26	64.46	75.84	54.10
0dB	47.74	53.90	61.50	47.58	52.68	44.67	55.86	58.51	56.56	53.90	28.49	39.69	34.09	49.45	36.60
-5dB	19.56	19.01	23.65	17.00	19.81	12.25	22.94	22.49	22.93	20.15	10.29	14.90	12.06	18.50	11.50
Average	80.97	84.66	85.53	80.38	82.89	80.84	83.89	84.65	83.54	83.23	71.25	76.60	73.93	81.23	
Improv.	35.63	74.59	54.65	42.96	51.96	68.60	55.79	73.52	66.05	65.99	23.82	-26.63	-1.41		46.90

Table 2. Clean training results on Aurora 2 database using FEAT1

6. ACKNOWLEDGEMENTS

This work was supported in part by NSF,¹ and by STM, HRL, and Broadcom and the state of California thru the UC Micro Program.

7. REFERENCES

- [1] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy condition," *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France*, 2000.
- [2] *ELRA 2001*.
- [3] Q. Zhu, M. Iseli, X. Cui, and A. Alwan, "Noise robust feature extraction for asr using the aurora 2 database," *Proc. of EuroSpeech*, pp. 185–188, 2001.
- [4] L. Lamel, L. Rabiner, Rosenberg A, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 29, pp. 777–785, 1981.
- [5] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," *Proc. of EuroSpeech*, pp. 1887–1890, 2001.
- [6] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," *Proc. of ICASSP*, pp. 1783–1786, 2000.
- [7] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 451–464, 1997.
- [8] Q. Zhu and A. Alwan, "Amplitude demodulation of speech and its application to noise robust speech recognition," *Proc. of ICSLP*, vol. 1, pp. 341–344, 2000.
- [9] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.
- [10] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 578–589, 1994.
- [11] *David Pearce, Personal communication*.

¹This material is based upon work supported by the National Science Foundation under Grant No. ANI-0085773. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).