Noise Robust Speech Recognition

Abstract

Speech recognition has seen extensive usage in commercial applications. Major challenges in speech recognition pertains to maintaining performance despite noise and individual distortions. The paper proposes a focus on clean signal enhancement and normalization during feature extraction to combat noise and formant differences from gender. Many techniques and combinations are proposed and analyzed on Aurora2 dataset. The best performance was achieved with vocal-tract normalization, variable frame-rate, silence normalization and other noise-robust techniques with an overall accuracy of 84.86%.

1. Introduction

Speech recognition is the task of interpreting speech signals to words. Speech recognition system is often applied to company answering machines, personal computer assistants, and automating note-taker.

The objective of this project is to design an automatic speech recognition (ASR) system front end. ASR front ends handle the brunt of signal processing by extracting raw data to extract useful and machine-interpretable features. A machine learning algorithm (the back-end) then takes in these features to estimate the words being uttered. Due to the variation of data in real application, the front end should minimize difference caused by effects such as noise, accents and genders. With these tenets in mind, we propose a front end that comprises a series of noise-robust processing stages and transform noisy data to an approximate cleaner version.

Humans can recognize speech in noisy background even when the speech is distorted by gender, age and accents. In fact, human perception still outperforms the state-of-the-art ASR by about a decade of error rate. In this paper, our front end confronts two main problems: noise and gender-based deviations. Due to the need to adapt to different environments arise in applications, the ASR training is limited to clean, male speech, but the ASR is evaluated on data from other environments. Differences in gender will reflect in different formant position, leading to misclassification. Noise will either cause the system to identify noise as speech or vice-versa.

Our front end performs a series of transformations to minimize the difference caused by noise and gender. We employ both old techniques – such as Peak Isolation, Variable Frame-Rate Analysis, Harmonic Demodulation, Pitch-based Vocal Track Normalization– as well as our new methods – such as Silence Normalization – to enhance Mel-Frequency Cepstral Coefficients.

This paper is organized as follows: Section 2 examines the problem setting; Section 3 presents a literature survey; Section 4 discusses the baseline performance; Section 5 describes in depth our front end; Section 6 evaluates the performance; Section 7 points to some future directions.

2. Problem Formulation

2.1. Assumptions

The project tackles the problem of implementing a speech recognizer front-end that is both noise-robust and gender-invariant. To approximate real data's variability, training is done on clean, male speech, but the system is evaluated on sets of male and female speech of varying noise levels. Our ASR works under the following assumptions:

1. The duration each word is relatively constant (the speakers do not attempt to talk excessively fast or slow).
2. The noise level is constant (there are no noiseless segments in noisy recordings) so that the short-time signal-to-noise ratio could be considered as constant.
3. Each recording contains at least one word.
4. The speech signals are flat (there is no yelling or singing during the production of speech signals) so the signal could be consider approximately stationary over short time segment.

2.2. Data Set

The data consists of male and female digit utterances under various babble noise with different SNR from the Aurora2 dataset. The utterances consist of random permutation of the numbers '0' to '9' and the sound 'oh' of varying length.

The training is done on 4220 recordings of clean male speech but the testing is done on the following sets: clean male speech, clean female speech, male speech with SNR 10 dB, female speech with SNR of 10 dB, male speech with SNR 5 dB, female speech with SNR 5 dB. The noise in the noisy recordings is babble noise– the sound of groups of people holding conversation in the background. The test data is in the same format as training data, and each set contains about 500 recordings.

For evaluation, a confusion matrix is generated, listing how each digits in the set was classified. The performance is evaluated using word accuracy defined by $\frac{\#\text{ correct} - \#\text{ inserted}}{\#\text{ digits}}$.

3. Literature Survey

While Mel-Frequency Cepstral Coefficients (MFCC) had been successful in noiseless speech recognition, it is well-known that the presence noise significantly degrades the performance of MFCC [1]. There has been many attempts to address MFCC's weakness with noise-compensation techniques [2], but recently a new set of features called Power-Normalized Cepstral Coefficents (PNCC) based on perceptual gamma tone filters have surpassed MFCC in performance [3]. Furthermore, with the rising popularity of deep learning, there is a trend in feeding the raw data to neural networks directly [4].

On the other hand, Vocal Tract Length Normalization (VTLN) based on frequency warping is also popular in speaker

normalization. There are many approaches to VTLN, some simply use the third formant F3 and some rely on subglottal resonance [5].

In this project, we take the approach of enhancing MFCC with noise-compensation/suppression techniques to device a set of noise-robust features. Furthermore, a pitch-based VTLN is used to eliminate the difference between male and female speech. Because the goal is to design a set of noise-robust features, the back end classifier is fixed as a Hidden Markov Model with Gaussian Mixture Model.

4. Baseline

4.1. Front End Features

The front end features are the first 13 MFCCs as well as the first and second derivatives for a total of 39 features per frame. The MFCCs are calculated using frames of length 25 ms and set size 10 ms. No noise-robust processing is applied.

4.2. Back End Classifier

The back end classifier uses left-to-right Hidden Markov Model and Gaussian Mixture Model for observation probability. There are 13 isolated word including '0' to '9', 'oh', silence and short pause. Every digit has 18 states; silence has 5 states; short pause has 3 states; all including the entry and exit state. Each state in the digit word model has 3 mixtures, each state in the silence and short pause models has 6 mixtures. Each Gaussian mixture of all the states has 39 dimensions.

4.3. Baseline Accuracy

Figure 1 shows the baseline accuracy of MFCC with HMM+GMM. Note that while MFCC performs well on male-clean data, the accuracy decays quickly as SNR decreases. At 5 dB, the accuracy on male speech is only 45.43%.

There is also significant degrades on with inter-gender performance. The classifier on clean female speech has only 82.72% compared to 99.44% on clean male speech. Furthermore, at 5dB, the accuracy has decayed to 12.35%.

	File Name	Training Data	Testing Data	MFCC Baseline
1	test_male_clean.m	male, clean	male, Clean	99.44 %
2	test_female_clean.m	male, clean	female, Clean	82.72 %
3	test_male_10dB.m	male, clean	male, 10dB-SNR babble noise	82.45 %
4	test_female_10dB.m	male, clean	female, 10dB-SNR babble noise	36.51 %
5	test_male_5dB.m	male, clean	male, 5 dB-SNR babble noise	45.43 %
5	test_female_5dB.m	male, clean	female, 5 dB-SNR babble noise	12.35 %

Figure 1: *Acuracy of baseline features.*

5. Noise Robust Enhancement

5.1. Overview

The feature used in this project is MFCC enhanced by a series of modules shown in Figure 2. First, the power spectrum is calculated from the raw speech. Then, it is passed through the harmonic demodulation and noise flooring block to remove the effect of glottal pulse and reduce the noise. Next, a silence-normalization block is used to reduce the effect of noise on non-speech segment. Then MFCCs are computed with 22 channels and a standard raised-sine lifter. Next, Variable Frame Rate Analysis is used to control the frame rate. We then apply Peak Isolation for more noise-compensation. Cepstran-mean subtraction is also applied to center the features. Finally, the first and second derivatives are computed from MFCC to be used as features with MFCC.

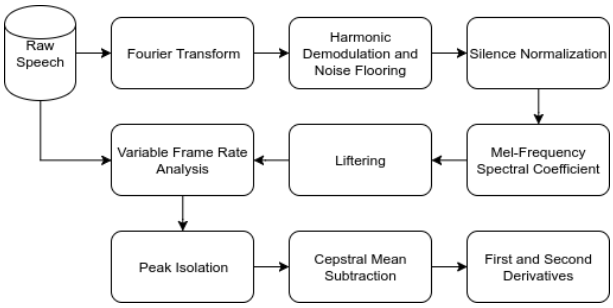

Figure 2: *An overview of the system.*

5.2. Voice Activity Detection (VAD)

The voice-activity detector (VAD) is a technique that classifies frames as either speech or non-speech. The goal is to isolate speech segments and control which noise processing algorithm is applied. Techniques used later on, such as pitch detection and silence normalization, are dependent on the VAD output for silence-speech distinguishing. For example, babble noise could contain valid harmonics and the pitch detector would give an incorrect pitch value as a result.

The VAD returns '1' or '0' for each frame, depending on whether speech is detected. The VAD first calculates the short-time energy (STE) in speech band for each frame, which are then binned as a histogram with 10 discrete bins. Figure 3 displays an example of such binning. Due to variation in volume and vowel, speech often varies more in energies than noise does. Therefore, the bin with the most values is assumed to be the noise. The noise mean and variance is then calculated from this bin and the VAD returns '1' if the energy in the frame exceeds three standard noise deviations from the noise mean. The VAD function is subsequently smoothed with median filter. This histogram-based VAD is highly noise-robust. A sample VAD output is displayed in Fig 4.

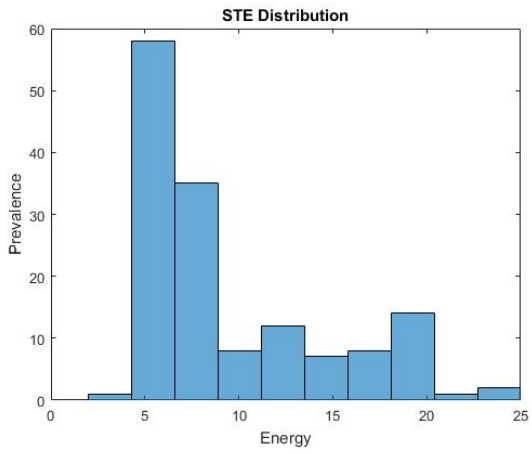

Figure 3: *Histogram of Short-Time Energy.*

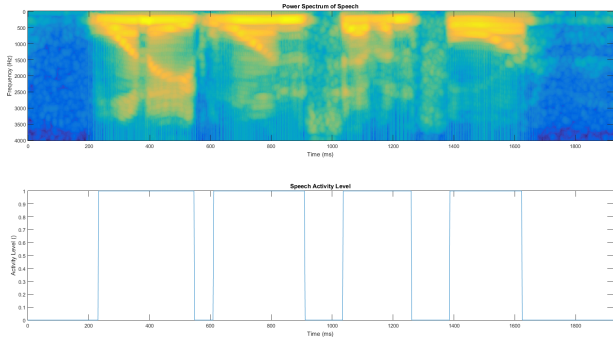

Figure 4: *Sample output from VAD with spectrogram.*

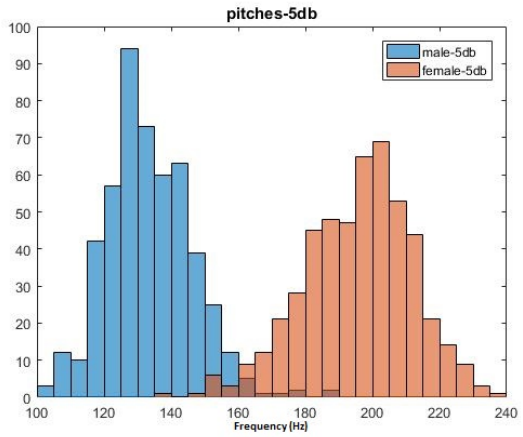

Figure 5: *Male and Female pitches of the 5db testing sets*

5.3. Pitch-based Vocal Tract Normalization (P-VTLN)

Vocal Track Length Normalization (VTLN) is a technique for normalizing gender disparities in vowel formants. The average male is taller than the average female and the difference in height is reflected vocal track length – males have a vocal tract 15% longer than females. Therefore, we expect the formants for male speakers to be lower than female formants. The goal of VTLN is to warp the frequency of the female speaker so that the female formants matches the male formants. We implement VTLN using a pitch-dependent, bilinear warping. The pitch is first estimated with a pitch detector. If the estimated pitch high, then each frame has its frequency warped according to a bilinear function represented in Fig 6.

The pitch detector estimates speaker's average pitch by finding the highest peak of the autocorrelation of the power spectrum. The autocorrelation's maximum is at zero, but its second highest maximum will occur at the autocorrelation lag corresponding to the fundamental frequency where all the harmonics align again. VTLN uses the VAD output to average the pitch of all speech frames.

5.4. Variable Frame Rate Analysis (VFR)

While speech signals are known to be time-varying and non-stationary, the signal characteristic changes at a slow rate. Thus speech processing literatures usually consider speech signal to be quasi-stationary over short time durations. Basing off this idea, most speech-analyzing algorithms work on a frame-by-

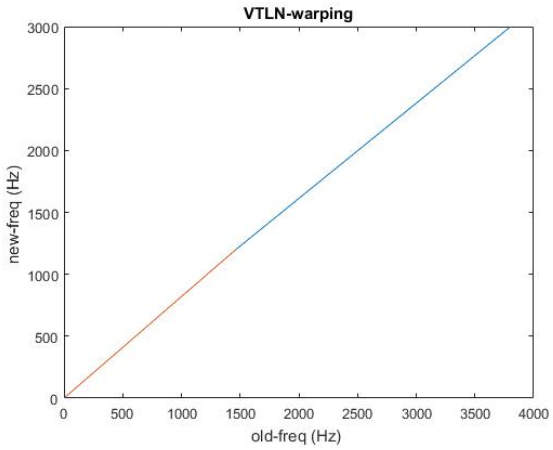

Figure 6: *VTLN Warping Function.*

frame basis – speech signals are windowed into frames with frame duration 20 ms to 30 ms and frame steps size 10 ms.

This kind of framing technique, however, assumes that the rate at which the signal characteristic changes remains constant. In practice, speech signal does not change at a constant rate – there is almost no change during the pronunciation of a vowel, but the signal changes rapidly in the phoneme transition region.

A naive way of addressing this issue is by over-sampling the frames with frame rate 2.5 ms. This unfortunately incurs a significant increase in the time for feature extraction and HMM model training. Variable Frame Rate Analysis offers a compromise by adapting the frame rate at different region of the signal.

In this project, we use a version of VFR described by You, Zhu and Alwan in [6] that uses entropy as the frame rate criterion. The intuition is that when there is more information(entropy), one should increase the frame rate to capture it.

Roughly speaking, entropy-based VFR first frames the speech signal with frame duration 25 ms and frame rate 2.5 ms. Then, each frame is taken as a realization of a Gaussian random vector. The local entropy of the signal is then approximated using a 30 ms neighborhood. Once local entropy is known, the frame rate is determined with simple thresholding: let $H(v_i)$ be the entropy at frame i, then we set the frame rate as

$$\text{frame rate } = \begin{cases} 5 \text{ ms} & \text{if } H(v_i) \geq T_1 \\ 7.5 \text{ ms} & \text{if } T_1 > H(v_i) \geq T_2 \\ 10 \text{ ms} & \text{if } T_2 > H(v_i) \geq T_3 \\ 12.5\text{ms} & \text{if } T_3 > H(v_i) \end{cases}$$

where T_1, T_2, T_3 are thresholds set by

$$\begin{aligned} T_1 &= w_1 M_{\max} + (1 - w_1) M_{\text{med}} \\ T_2 &= (1 - w_2) M_{\max} + w_2 M_{\text{med}} \\ T_3 &= (1 - w_3) M_{\text{med}} + w_3 M_{\min} \end{aligned}$$

and $M_{\max}, M_{\text{med}}, M_{\min}$ are the maximum, median, and minimum of the relative entropies across the whole signal.

Figure 7 shows an example of VFR on male voice with 5 dB of babble noise. Note that the speech region roughly corresponds to region with higher frame rate. This example illustrates that entropy-based VFR is robust to noise and thus is suitable for our purpose.

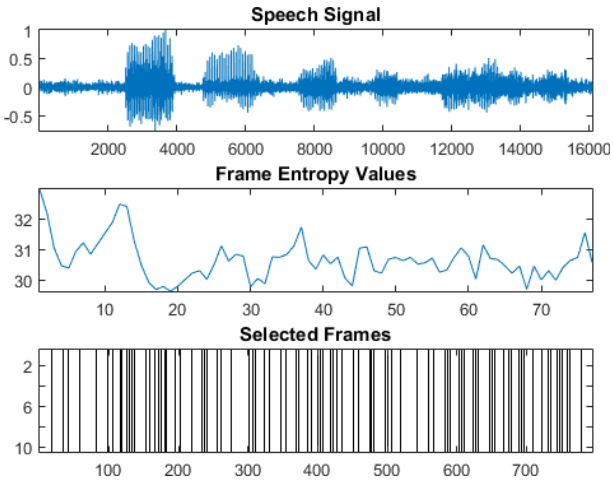

Figure 7: *An example of entropy-based VFR with male voice at 5 dB babble noise.*

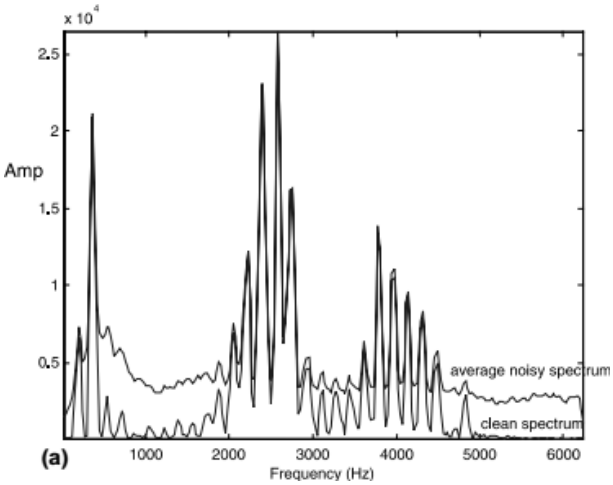

Figure 8: *Speech power spectrum with and without noise.*

5.5. Harmonic Demodulation (HD)

In speech processing, speech signal is typically modeled as the output of glottal pulse passing through the vocal tract transfer function. Since the glottal pulse is mostly constant, it carries little information. Conversely, most of the information is contained in the vocal tract function. Because convolution is time domain is multiplication in frequency domain, one can consider this as modulation of message (vocal tract transfer function) by the carrier (glottal pulse) in frequency domain. In analogy to communication, one approach to recover speech information is by performing demodulation to recover the vocal tract transfer function. Suppose $S(n)$ is the modulated wave (speech power spectrum), then demodulation can be done by convolving with a low pass filter $h(n)$ in frequency domain. i.e., we have

$$\hat{M}(n) = (S * h)(n)$$

where $\hat{M}$ is the estimated message (VTTF).

This traditional approach with linear filter, however, is not noise robust. Suppose the speech is corrupted by an additive noise, then in [7], Zhu and Alan showed that, on average, spectral region with low power is most affected by noise. Figure 8 shows an example of clean speech spectrum and a corrupted version of it. Note that the harmonic peaks are virtually unchanged by noise but the harmonic valleys are corrupted.

Since convolution does not discriminate points based on its magnitude, the noisy low-power part can corrupt the result of demodulation. We address this problem with a non-linear filter:

$$\hat{M}(n) = \max_k S(k)h(n-k)$$

This filter reduces the effect by noisy harmonic valleys by focusing on the peak, which leads to a more noise-robust system.

5.6. Noise Flooring (NF)

The aforementioned non-linear harmonic demodulation filter helps suppressing the noise a harmonic valleys. Unfortunately, it does little to combat the noise at formant valleys, so an additional step is needed to suppress or normalize them. In the same paper [7], Zhu and Alwan also proposed a technique called noise flooring to address this issue. In noise flooring, all power spectrum values whose level is below a certain threshold is set to a constant level. This normalizes the power at formant valleys, thus reducing the effect of noise. For this project, the threshold is set to 0.4 of the average power.

5.7. Peak Isolation (PKISO)

Peak Isolation(PKISO) is a noise-robust technique that reduces the impact of noises. PKISO first converts the liftered MFCC back into the mel-frequency spectrum with IDCT, then rectifies the spectrum (setting all values less than zero to zero). The resultant spectrum is then converted back into cepstrum coefficients via DCT to obtain the peak-isolated MFCCs.

PKISO works under the notion that noise affects the spectrum valleys much more than the spectrum peaks. However, information of the vowels resides in the spectrum peaks, so the distortion in the spectrum valleys will contribute to back-end susceptibility to noise. PKISO removes spectrum valleys via rectification, reducing the impact of noise on the resultant feature. Fig 8 showcases the effect noise has on the spectrum valleys. While the spectrum peaks align for noisy and clean signals, the spectrum valley deviates significantly between the two. By setting a threshold at 0, the spectrum valley is eliminated.

5.8. Silence Normalization

Silence Normalization (SN) is a noise-robust technique we introduce in this paper. Because the training is done on clean data, the classifier often mistakes the babble noise in silence region as speech. SN alleviates this by damping the noisy silence region.

For each recording, SN first detects speech-activity level with VAD, then normalizes the energy of the non-speech segments across all recordings. The result of SN is that the training and testing data all have roughly the same signal-to-noise ratio and the influence of noise on the data is reduced. SN is reliant an estimated SNR to control the level of non-speech damping.

The signal-to-noise ratio (SNR) is a measurement of the relative energy of the clean speech to the noise energy. The SNR is assumed to be constant across all recordings in a set, but its value is not known by the front end. Our algorithm estimates the SNR using the VAD as the follows: first we calculates average noise energy from non-speech segments; then signal energy is estimated by subtracting the average speech energy from the average noise energy. The SNR is then calculated using:

$$SNR = 10 \log \left(\frac{\widetilde{P}_{\text{clean}}}{\widetilde{P}_{\text{noise}}} \right)$$

Figure 9 shows the SNR estimations on male and female sets. While there is significant overlap between the noisy sets, the clean and noisy sets are nicely divided.

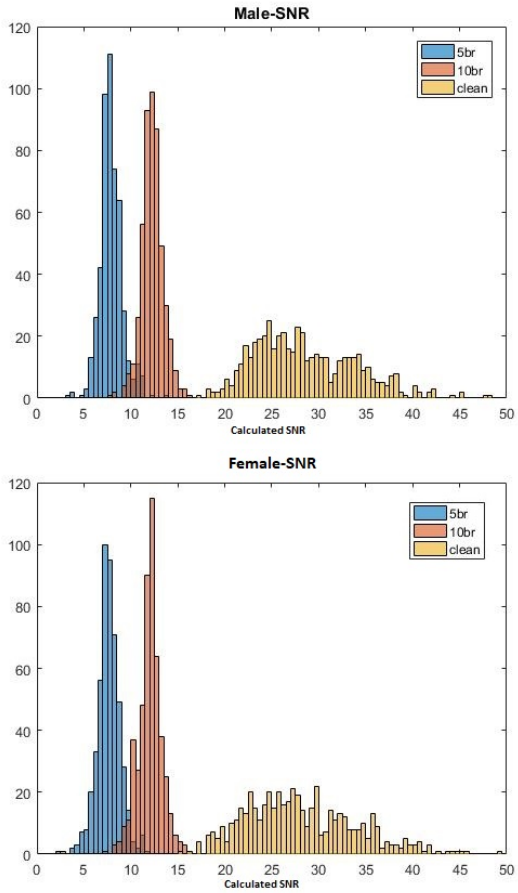

Figure 9: *SNR estimation of 6 testing data sets*

SN does not affect frames where speech is detected, but it multiplies frames of non-speech with a normalizing factor. The non-speech factor varies from one to a normalization constant dependent on the frame's proximity to a speech frame. This minimizes distortion to the fricatives, which often surround the vowels and are not present in the VAD. The non-speech factor is given by a sigmoid (x is the distance to a speech frame):

$$\text{factor} = \frac{(1 - \text{NormConst})}{1 + e^{-x-2}} + \text{NormConst}$$

The normalization constant is determined by the recording's SNR. Signals with high SNRs receives a normalization constant greater than one, which enhances the noise; signals with low SNRs receives a constant less than one, which dampens the noise. Figure 10 showcases the SN factors for a particular recording. In our implementation, we used a normalization constant of 1 if the SNR exceeded 16.5 dB and 0.01 otherwise. Fig 11 demonstrates before SN and after SN spectrograms of a male 5db recording.

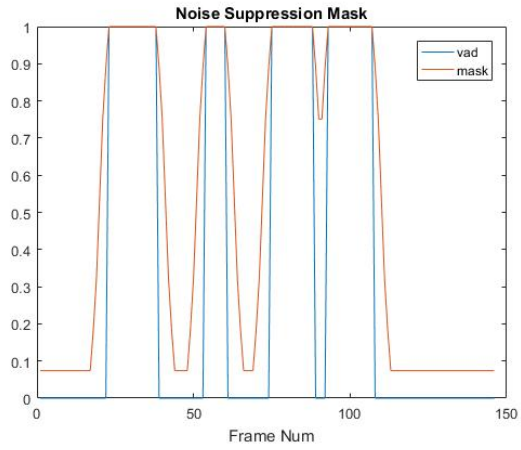

Figure 10: *Per-frame Damping Factor and VAD for a 5db male recording*

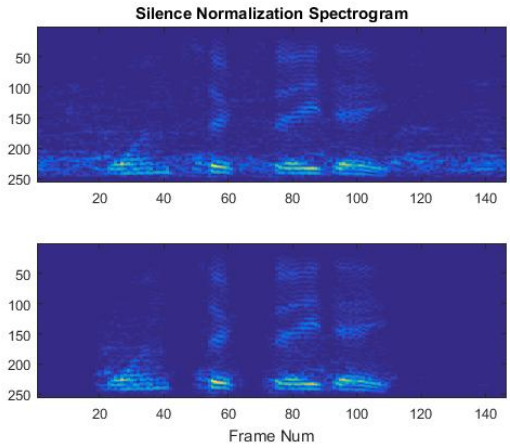

Figure 11: *Comparison of Spectrograms before-SN (Top) and after-SN (Bottom)*

6. Results and Discussion

6.1. Performance

Table reftbl:accuracy shows the word accuracies of different combination of noise-robust modifications on the six testing sets. Only the best performance for a combination is reported since modifications have several parameters to tune.

The ASR achieves the highest accuracy with the front end implementing vocal-tract length normalization (VTLN), variable frame-rate (VFR), silence normalization (SN), peak isolation (PI), noise flooring (NF), and harmonic demodulation (HD) with an total average accuracy of 84.4% across all testing sets. Baseline implementations, such as cepstral mean subtraction, keeping zeroth cepstrum coefficient, first and second order cepstrum coefficients, are also included.

The results indicate that including PI incurs heavy trade-off on clean data. In the case of VTLN+SN+PI we see a 5-10% decrease in clean performance and 5-10% increase in 5 dB noise accuracy. This effect can still be observed in the best case(VTLN+SN+PI+VFR+HD), where the clean male and female are still 3-6% lower than their peak accuracy. The reason this trade-off occurs is unknown. One potential explanation is

feature	clean male	clean female	10db male	10db female	5db male	5db female
base	99.44	83.95	83.2	31.5	51.21	18.87
VTLN	99.82	98.82	82.89	69.19	52.15	39.21
NS	99.38	84.30	89.42	49.74	71.81	34.51
PI	95.21	61.08	72.12	37.33	54.53	20.4
VTLN +SN	99.44	98.82	89.23	77.43	71.87	55.79
VTLN +SN+PI	91.95	88.46	87.75	80.88	76.34	65.31
VFR	98.26	67.14	88.61	43.56	63.53	24.51
VFR +HD	98.32	66.61	86.93	42.68	61.42	25.46
vFR +HD+NF	98.38	67.31	86.50	41.92	60.73	24.22
VFR +HD+NF +SN+PI	96.58	60.32	92.41	57.14	78.22	41.68
VTLN +VFR +HD+NF +SN+PI	96.58	92.06	92.22	85.19	78.03	65.08

Table 1: *Word accuracy of different feature set under various conditions*

that the removal of spectral valleys ends up taking away vowel identity information. Another explanation is that PI unintentionally amplifies noise if the short-time noise is loud enough. Another interesting attribute about PI is the large mismatch between average word correctness (72.41%) and average word accuracy(58.3%). This implies that PI in the front end causes the back end to become particularly susceptible to insertions.

SN improves the noise performance drastically with a suppression level of 0.01. But setting the suppression level to zero leads to a performance of 10-15% worse than baseline.

6.2. Error Evaluation

Figure 12 shows the confusion matrix for male speech with 5 dB noise with our best feature set. In Figure 12, we see that the most probable errors are confusing '4' with '1', '2', '6', and 'oh'. Another probable source of error comes is '8' with '6'. Here, the general trend is that words with the pattern fricative-vowel-fricative are confused with other words following the same pattern (the fricatives can be omitted). There is a possible explanation for this. One of our key modules performs silence-normalization by suppressing the part of signal classified as non-speech by the VAD. Due to the fact that the VAD is not perfect, the act of suppression might have attenuated the fricatives so much that the classifier registers them as silence, leading to errors like '4' to 'oh'.

7. Conclusions

In this project, we proposed an enhanced MFCC for noise-robust automatic speech recognition that can also handle mismatch due to gender difference. This set of features is created by combining MFCC with Harmonic Demodulation, Silence Normalization, Variable Frame Rate Analysis, and Peak Isolation. When tested on Aurora2 dataset, we find that this enhanced MFCC suffers a modest degrade in performance on

```
------------------------ Overall Results ------------------------
SENT: %Correct=52.30 [H=262, S=239, N=501]
WORD: %Corr=87.24, Acc=78.10 [H=1402, D=56, S=149, I=147, N=1607]
------------------------ Confusion Matrix ------------------------
       O   T   T   F   F   S   S   E   N   Z   O
       n   w   h   o   i   i   e   i   i   e   h
       e   o   r   u   v   x   v   g   n   r
               e   r   e       e   h   e   o
               e               n   t               Del [ %c / %e]
 One  140   4   0   0   0   0   0   1   2   0   5    5 [92.1/0.7]
 Two    0 120   1   0   0   4   0   4   2   1   2   10 [89.6/0.9]
Thre    1   7 137   0   0   3   2   2   0   2   0    2 [89.0/1.1]
Four    7   7   1 101   0   5   1   2   0   0  12    6 [74.3/2.2]
Five    0   0   0   0 143   1   0   0   5   0   1    2 [95.3/0.4]
 Six    0   9   0   0   0 108   1   1   0   0   2    5 [89.3/0.8]
Seve    0   0   0   0   0   6 133   1   0   0   0    3 [95.0/0.4]
Eigh    0   3   1   0   1  11   0 122   2   0   2   11 [85.9/1.2]
Nine    0   1   0   0   2   0   1   0 131   0   0    3 [97.0/0.2]
Zero    0   6   0   0   0   0   0   0   2 125   2    1 [92.6/0.6]
  Oh    0   1   1   4   0   0   2   1   0   1 142    8 [93.4/0.6]
 sil    0   0   0   0   0   0   0   0   0   0   0    0
Ins     9  16   1   4   5  32   6  25  11   1  37
```

Figure 12: *Confusion matrix for male speech with 5 dB noise.*

clean male data (from 99.44% to 96.58%). However, our feature set behaves significantly better than pure MFCC on data with gender mismatch or noise: on clean female data, the accuracy improves from 83.95% to 92.06%; on 5 dB male, from 51.21% to 78.03%; on 5 dB female, from 18.87% to 65.08%.

As discussed in 6.2, a large portion of error can be attributed to the fricatives being attenuated by the silence-normalization module. Thus, one way to improve this system is by implementing a better VAD. A less aggressive silence-normalization can also be attempted.

Another approach to improve the system is by extending this type of modification to other features such as LPCC, PLP, and PNCC. How feature enhancement interacts with features that are already robust to noise such as PNCC is unknown, but it could lead to a much better performance than what is obtained by pure PNCC and our enhanced MFCC.

8. References

[1] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[3] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4101–4104.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[5] H. Arsikere and A. Alwan, "Frequency warping using subglottal resonances: Complementarity with vtln and robustness to additive noise," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6299–6303, 2014.

[6] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to asr," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. I–549–52 vol.1.

[7] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer Speech and Language*, vol. 17, no. 4, pp. 381 – 402, 2003.