PROBABILISTIC APPROACH FOR SPEECH INTELLIGIBILITY IMPROVEMENT AND NOISE REDUCTION

D. Sreekanth,
Abhinav Hi-Tech College of Engieering, Hyderabad
dk_sreekanth@yahoo.com

D. Sunitha
Abhinav Hi-Tech College of Engieering, Hyderabad
sunithadasari1@gmail.com

ABSTRACT

In speech processing applications often it is observed that many algorithms implemented so far in the past were able to concentrate either on reducing the noise or improving the speech intelligibility, but not the both. The algorithm introduced in this paper focuses on reducing the noise in the speech signal while improving its intelligibility. The new algorithm is based on probabilistic synthesis and analysis of speech signal.

Key words: speech intelligibility, synthesis, baysian probability, binary mask.

1. INTRODUCTION

Unlike the most speech enhancement algorithms improve speech quality, they may not improve speech intelligibility in noise, this work focuses on the development of an algorithm that can be optimized for a specific acoustic environment and improve speech intelligibility. Ideal binary time-frequency masking is a signal separation technique that retains mixture energy in time-frequency units where local signal-to-noise ratio exceeds a certain threshold and rejects mixture energy in other time-frequency units.

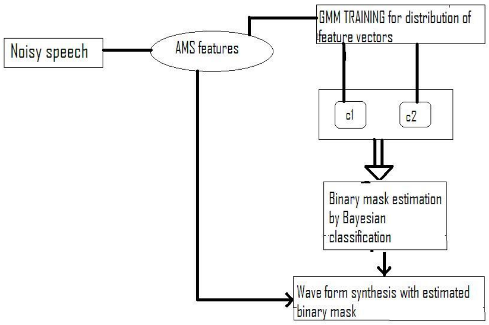
Noisy speech
AMS features
GMM TRAINING for distribution of feature vectors
c1
c2
Binary mask estimation by Bayesian classification
Wave form synthesis with estimated binary mask

Figure 1.1 General Block diagram of the algorithm

We improve intelligibility of speech synthesized via an algorithm that decomposes the input signal into T-F regions, with the use of a crude auditory-like filter bank, and uses a simple binary Bayesian classifier to retain target-dominated spectro-temporal regions while removing masker-dominated spectro-temporal regions. Amplitude modulation spectrograms (AMSs) are used as features for training Gaussian mixture models (GMMs) to be used as classifiers. Figure1.1 shows the general block diagram of algorithm. In noisy environments, the speech signal SNR is very low and negative some times. Algorithms that improve speech quality do not necessarily improve speech intelligibility. This is most likely due to the distortions introduced to the speech signal. In contrast to speech quality, intelligibility relates to the understanding of the underlying message or content of the spoken words, and is often measured by counting the number of words identified correctly by human listeners. Intelligibility can potentially be improved only by suppressing the background noise without distorting the underlying target speech signal.

2. BASIC DEFINTIONS

2.1 Speech Segmentation or Framing:

In speech processing it is often advantageous to divide the signal into frames to achieve stationarity. Normally a speech signal is not stationary, but seen from a short-time point of view it is assumed as stationary during 10-30ms. Framing is used to cut the long-time speech signal into short time signals in order to get stationarity or stable frequency characteristics. The time for which the signal is considered for processing is called a window and data acquired in a window is called as a frame. Generally frames are over lapped in order to get typical feature in that duration.

2.2 Sub band filtering

In signal processing, an incoming signal is decomposed into different frequency bands or channels, is usually done by using a collection of filters called Filter Bank. A filter bank is an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency subband of the original signal. One application of a filter bank is a graphic equalizer, which can attenuate the components differently and recombine them into a modified version of the original signal. The process of decomposition performed by the filter bank is called analysis (meaning analysis of the signal in terms of its components in each sub-band); the output of analysis is referred to as a subband signal with as many subbands as there are filters in the filter bank. The reconstruction process is called synthesis, meaning reconstitution of a complete signal resulting from the filtering process.

2.3 Features of speech signal

Feature is a distinctive characteristic of a speech unit that serves to distinguish it from other units of the same kind. These features play a vital role in many speech processing applications like enhancement, compression and especially in speech recognition.

There are different types of features such as Real Cepstral Coefficients (RCC), Mel Frequency Cepstral Coefficients (MFCC), Delta Mel frequency Cepstral Coefficients (ΔMFCC), Delta Delta Mel Frequency Cepstral Coefficients (ΔΔMFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Amplitude Modulation Spectrograms (AMSs).

2.4 Spectrogram:

The spectrogram is the plot estimate of the short-term (time) frequency content of the signals in which a two dimensional representation of the speech intensity. It is a time Vs frequency plot. Mathematically, the spectrogram of a speech signal is the magnitude square of the Short Time Fourier Transform of that signal.

2.5 Gaussian Mixture models

Mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population.

While problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population-identity information.

Speech production is not deterministic in a particular sound (e.g., a phone) is never produced by a speaker with same vocal tract shape and glottal flow, due to context coarticulation, and anatomical and fluid dynamical variations. One way to represent this variability is probabilistically through a multi dimensional Gaussian pdf. The Gaussian pdf is a state-dependent in that there is assigned a different Gaussian pdf for each acoustic sound class. The states are like Quasi periodic, noise-like, and impulse like sounds or on a very fine level such as individual phonemes.

2.6 Binary Mask

In noisy environments, the speech signal SNR is very low and negative some times. Algorithms that improve speech quality do not necessarily improve speech intelligibility. This is most likely due to the distortions introduced to the speech signal. In contrast to speech quality, intelligibility relates to the understanding of the underlying message or content of the spoken words, and is often measured by counting the number of words identified correctly by human listeners. Intelligibility can potentially be improved only by suppressing the background noise without distorting the underlying target speech signal. The pursued approach is motivated by intelligibility studies of speech synthesized using the ideal binary mask (IdBM) [15]–[17], which in turn requires access to the SNR at each frequency bin. The ideal binary mask (originally known as *a priori* mask [18]) is a technique explored in computational auditory scene analysis (CASA) that retains the time-frequency (T-F) regions of the target signal that are stronger (i.e., SNR>0 dB) than the interfering noise (masker), and removes the regions that are weaker than the interfering noise (i.e., SNR>0 dB). Previous studies have shown that multiplying the ideal binary mask with the noise-masked signal can yield large gains in intelligibility, even at extremely low (-5, -10) dB SNR levels. In these studies, prior knowledge of the true spectral SNR and subsequently the ideal binary mask was assumed. In practice, the binary mask needs to be estimated from the corrupted signal requiring an accurate estimate (and classification) of the spectral SNR. This algorithm decomposes the input signal into T-F units with the use of a crude auditory-like filter bank and uses a simple binary Bayesian classifier to retain target-dominant1 T-F units while removing masker-dominant units. Amplitude modulation spectrograms (AMS) were used as features for training Gaussian mixture models (GMMs) to be used as classifiers. Unlike most speech enhancement algorithms, the proposed algorithm did not require speech/noise detection nor the estimation of noise statistics.

3. PROBABILISTIC ALGORITHM

3.1 Feature Extraction

1. Here the clean speech ‘cl’ and noisy speech ‘ns’ both of duration 2 seconds is taken. The sampling rate is 12 kHz.
2. Now we calculate the envelope for each 0.25ms duration, that is 3 samples are considered as a single sample or decimated by a factor of 3 for AMS feature extraction. Envelope is simply the absolute value of the samples.
3. Framing is done with a duration of 32ms (that is 384 samples for actual speech and 128 samples for AMS frame).
4. Number of frames is given by [(length of speech)/(no. of samples in 32ms frame)- (no. of samples in 32ms frame)/ (no. of samples in overlap step)+1]=162 frames.

5. The speech frequency is sub band filtered in to 25 bands.

6. As mel frequency spacing is used, a mel filter bank is designed.

7. Here low frequency=0Hz and high frequency=sampling rate/2 are taken.

8. Now get the low, high, centre mel frequencies as melfreq=1000*log(1+f/800)center mel_freq=low+(1 to 26)*(high-low/26)10. Now lower cutoff frequencies=centers(1 to 25)Higher cutoff frequencies =centers(2 to 26)Centers=lower cutoff + higher cutoff frequency/2.

9. Band widths are calculated and time indexes are calculated as freq/Samplingrate*no.of samples in the 32ms frame.

10. Now butter worth low and high pass filters are designed

11. Signal of a subband in time domain is obtained with this designed filter coefficients.

12. Now framing part is done by simple logic. Where each frame has 128 samples multiplied with a hanning window. 128 zeros are padded and a 256 point FFT is calculated. Each framed speech segment in a sub band is called a TF unit.

13. Similar to that of mel filter bank, uniformly spaced 15 triangular windows are designed in that FFT spectrum. These 15 windows are multiplied and summed up to get 15 dimensional AMS feature vector.

14. Like this it is done for each and every TF unit. Thus feature extraction part is implemented. Here for 25 bands, 162 time frames, a total of 4050(25*162) TF units and 60,750(25*162*15) features or data points are obtained.

15. Signal to noise ratio (SNR) in TF unit is obtained by SNR = 10*log10[(clean)./(true noise)] where true noise=noisy speech-clean speech

3.2 Training of GMMs

1. These obtained features are used for training a Gaussian Mixture Model. These features are divided into two groups according to their SNR values.

2. This is done by grouping all the features whose SNR value is greater than -8dB as a first group and others as second group for the first 15 subbands. Similarly for the other 10 subbands the dividing threshold is set as -16dB.

3. For faster convergence of GMM each group is further divided in to 2 sub groups. So SNR thresholds of -4dB is set for lower bands and -10 dB is set for higher sub-bands. Each group is used for training a Gaussian pdfs. Likewise four pdfs are fitted to the four groups of features.

4. Probability of each group so called class is calculated as number of features in that class/total number of features.For the above steps GMMBAYES functions available in the MATLAB tool box is used.

3.3 Enhancement

1. Once GMMs are trained with training data, the binary mask is now implemented. A Bayesian classifier is used to estimate the binary mask.

2. As the aprior probabilities and Gaussian pdfs are known, Aposteriori probabilities that is the probability that a feature belongs to a class or group.

3. Here the first two groups and next two groups' Aposteriori probabilities are combined. The binary mask is now implemented by comparing two Aposteriori probabilities.

4. Now by simply multiplying the mask with the speech signal TF units that is the noise dominated TF units are eliminated. And the retained TF units finally mixed to get improved speech. Mixed in the sense the time frames are multiplied by hanning window and appended and the frequency bands are simply added.

5. Now for the same babble (speech shaped) noise and different speaker, the features are again classified by the trained GMMs and new binary mask is implemented. As it can be seen that there is no calculation of SNR's is required.

6. Similarly for a new type of noise again GMMs are trained and intelligibility is improved.

4. SIMULATION RESULTS

Here some of the sentences are taken and processed for intelligibility improvement. First clean speech is recorded without any background noise and then the speaker is asked to repeat the words in noisy environment where the noise is high i.e., SNR is very low. Such noisy speech is recorded and both the signals are given for GMM training and a binary mask is designed. Once the binary mask is designed the noisy speech is multiplied with the mask and synthesized. Now the speaker is asked to speak another sentence and again the new noisy speech is given to trained GMM to get new binary mask. This mask is multiplied with new noisy speech and processed to get improved intelligible speech. An example speech sentence is taken as “shake the dust from your shoes stranger” Its duration is 2sec and has the length of 31459 samples that is sampling rate is 12000 samples/sec. Then signal is divided into 25 bands according to mel frequency scale. The Mel filter bank is implemented as described above.

The mean performance, computed in terms of percentage of words identified correctly by the NH listeners, for sentences produced by male. A substantial improvement in intelligibility was obtained with the proposed algorithm using GMM models, compared to that attained by human listeners with unprocessed (corrupted) speech. The improvement was more evident at -5 dB SNR levels for all three maskers tested.

To quantify the accuracy of the binary Bayesian classifier, the average hit (HIT) and false alarm (FA) rates for three test sets is computed. Each test set comprised of 3 sentences, for a total of 9 sentences corresponding to 36,450 T-F units (162 frames 25 frequency bands) for the male-speaker sentences.

Performance comparison, in terms of hits (retained TF units) and false (eliminated TF units) alarm rates, of the AMS feature vectors for the male-speaker data at -5 dB SNR.

TF units	Babble	Factory	Speech shaped
Retained	79.4%	60.58%	76.12%
Eliminated	20.6%	39.42%	25.88%

Table 4.1: Comparison of retained speech signal

The noisy speech signal is processed by multiplying with a binary mask. Here the TF units of the noisy signal whose SNR less than -5dB are eliminated by the binary mask. The improvement (over 60% points in some cases) was more evident at −5 dB SNR levels for all three maskers tested.

Mean speech recognition scores obtained by 15 NH listeners for corrupted (unprocessed) sentences (denoted as UN) and sentences processed using the IdBM in the various SNR/masker conditions (-5dB and 0dB). Error bars indicate standard errors of the mean.

Only 20%, 3 out of 15 got correct perception at unprocessed noisy speech and after processing with the proposed algorithm over 90%, 11 out of 15 got correct perceptions of the spoken words when the SNR is -5dB for babble noisy speech which is a negative SNR implies extremely noisy condition. And when the SNR is at 0dB, 10 out of 15 recognized correct words before processing and 12 out of 15 recognized after processing with the proposed algorithm.

Similarly tests had been done for factory noise and speech shaped noise (random noise), and found that there is substantial increase in intelligibility scores, that is the number of persons who recognized correct words after processing with the proposed algorithm are considerably more than that of those who recognized before processing.

Percent of persons (out of 15), who correctly recognized the spoken words under different noisy conditions (babble, factory and speech shaped noise) and different SNR levels (-5dB and 0dB).

	Babble noise		Factory noise		Speech shaped noise	
	-5dB	0 dB	-5dB	0 dB	-5dB	0 dB
Un processed	18%	75%	40%	80%	45%	85%
After processing	92%	95%	90%	95%	92%	95%

Table 4.2 Percentage of intelligibilities (Percentage of persons who correctly identified) for various noisy environments and at different SNR levels before and after processing.

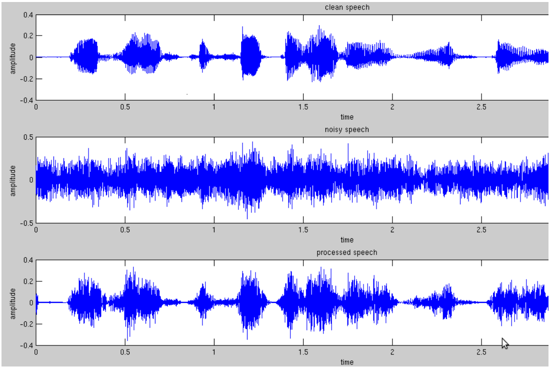

Figure 4.1: Clean, noisy and processed speech signal

5.CONCLUSION AND FUTURE SCOPE

Large gains in intelligibility were achieved with the proposed algorithm. The intelligibility of speech processed by the proposed algorithm was substantially higher than that achieved by human listeners listening to unprocessed (corrupted) speech, particularly at extremely low SNR levels (−5 dB). Attribute this to the accurate classification of T-F units into target- and masker-dominated T-F units, and subsequently reliable estimation of the binary mask. As demonstrated by several intelligibility studies with NH listeners, access to reliable estimates of the binary mask can yield substantial gains in intelligibility. The accurate classification of T-F units into target- and masker-dominated T-F units was accomplished with the use of neurophysiologically-motivated features (AMS) and carefully designed Bayesian classifiers (GMMs). Unlike the mel-frequency cepstrum coefficients features commonly used in ASR, the AMS features capture information about amplitude and frequency modulations, known to be critically important for speech recognition.

GMMs are known to accurately represent a large class of feature distributions, and as classifiers, GMMs have been used successfully in several applications and, in particular speaker recognition. Other classifiers (e.g., neural networks, and support vector machines) could alternatively be used.

A smaller number (25) of channels was used in this work for two reasons: (a) to keep the feature dimensionality small and (b) to make it appropriate for hearing aid and cochlear implant applications, wherein the signal is typically processed through a small number of channels.

The proposed algorithm can be used not only for robust ASR or cell phone applications but also for hearing aids or cochlear implant devices. Modern hearing aids use sound classification algorithms to identify different listening situations and adjust accordingly hearing aid processing parameters.

All advantages cited before the proposed approach suitable for trainable hearing aids and cochlear implant devices. As these devices are powered by a digital signal processor chip, the training can take place at the command of the user whenever in a new listening environment.

Following the training stage, the user can initiate the proposed algorithm to enhance speech intelligibility in extremely noisy environments (e.g., restaurants). However, a user might encounter a new type of noise not included in the training set. In such circumstances, either new training needs to be initiated or perhaps adaptation techniques can be used to adapt the parameters of existing GMM models to the new data.

REFERENCES

1. Brungart D., Chang P., Simpson B., and Wang D. (2006). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*120, 4007–4018. doi: 10.1121/1.2363929.

2. Dempster A. P., Laird N. M., and Rubin D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B (Methodol.)* 39, 1–38.

3. Furui S. (1986). "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.* ASSP-34, 52–59. doi: 10.1109/TASSP.1986.1164788

4. Hu Y., and Loizou P. C. (2007a). "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.* 122, 1777–1786. doi: 10.1121/1.2766778

5. Hu Y., and Loizou P. C. (2008). "Techniques for estimating the ideal binary mask," in The 11th International Workshop on Acoustic Echo and Noise Control, Seattle, WA.

6. Kollmeier B., and Koch R. (1994). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.* 95, 1593–1602. doi: 10.1121/1.408546

7. Li N., and Loizou P. C. (2008a). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J. Acoust. Soc. Am.* 123, EL59–EL64. doi: 10.1121/1.2884086

8. Li N., and Loizou P. C. (2008b). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.* 123, 1673–1682. doi: 10.1121/1.2832617

9. Loizou P. C. (2007). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL:).

10. Reynolds D., Quatieri T., and Dunn R. (2000). "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.* 10, 19–41. doi: 10.1006/dspr.1999.0361.

11. Zeng F.-G., Nie K., Stickney G. S., Kong Y.-Y., Vongphoe M., Bhargave A., Wei C., and Cao K. (2005). "Speech recognition with amplitude and frequency modulations," *Proc. Natl. Acad. Sci. U.S.A.* 102, 2293–2298. doi: 10.1073/pnas.0406460102.

12. Zakis J. A., Dillon H., and McDermott H. J. (2007). "The design and evaluation of a hearing aid with trainable