International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Optimizing Point of Care Interaction with Speech Recognition to Preserve Learnability and Intelligibility

M Narasimha Rao[1], A Venkata Raju[2], Matta Venkata Durga Pavan Kumar[3]

Submitted: 05/09/2024 **Revised:** 18/10/2024 **Accepted:** 26/10/2024

Abstract— Although the majority of speech augmentation algorithms increase voice quality, they may not augment speech intelligibility in noisy environments.

This work examines the creation of an algorithm that may be tailored to a particular acoustic environment to enhance speech intelligibility. The suggested technique disaggregates the input signal into time-frequency (T-F) units and employs a Bayesian classifier to make binary determinations on whether each T-F unit is predominated by the target signal or the noise masker. Target-dominated time-frequency units are preserved, but masker-dominated time-frequency units are eliminated. The Bayesian classifier is trained for each acoustic environment using an incremental method that perpetually adjusts the model parameters as further data is acquired.

Listening tests were performed to evaluate the intelligibility of speech synthesized using incrementally modified models based on the quantity of training sentences. The results demonstrated significant improvements in intelligibility, exceeding 60% in babbling at a 5 dB signal-to-noise ratio, with a minimum of 10 training phrases in babble and at least 80 words in other loud environments.

Index Terms—*Environment-optimized algorithms, speech enhancement, speech intelligibility.*

INTRODUCTION

Significant progress has been achieved in the creation of enhancement algorithms that may mitigate background noise and increase voice quality [1]. Significantly less advancement has been achieved in the development of algorithms aimed at enhancing voice intelligibility. As shown in [2], algorithms that enhance voice quality do not inherently enhance speech intelligibility. This is probably attributable to the distortions imposed on the voice stream. Unlike speech quality, intelligibility pertains to the comprehension of the fundamental meaning or substance of uttered words, often assessed by tallying the number of words accurately recognized by human listeners. Intelligibility may be enhanced just by mitigating background noise without altering the fundamental target voice signal. Algorithms designed to enhance voice intelligibility in loud circumstances would be very beneficial not just for mobile phone apps but also for hearing aids and cochlear implants. The creation of such algorithms has proven difficult for several decades, likely owing to the pursuit of algorithms capable of functioning across all forms of maskers (noise) and varying signal-to-noise ratio (SNR) levels, which is clearly an ambitious objective.

In some speech recognition applications (e.g., voice dictation) and hearing aid applications (e.g., [4]), the algorithm may be contingent upon the speaker and/or the surroundings.

Numerous environment-dependent methods have been lately proposed in references [5]–[10]. The originality of these algorithms is in the creation of spectral weighting functions (gain functions) that have been trained using a data-driven approach based on diverse error criteria. In contrast to the gain functions obtained for minimal mean square error (MMSE) and maximum a posteriori (MAP) estimators [11]–[13], the gain functions presented in [7]–[10] do not presume any specific probability density functions (pdf) for the complicated clean and noise spectra.

Fingscheidt et al. [10] used an extensive corpus of clean speech and noise data to develop frequency-specific gain functions for a particular noise environment. The gain functions were articulated as

[1,2] *Associate Professor, Department of Computer Science and Engineering, International School of Technology and Sciences for Women, A.P, India.*

[3] *Assistant Professor, International School of Technology and Sciences for Women), A.P, India.*

a function of the a posteriori and a priori signal-to-noise ratios (SNRs), calculated using a modified decision-directed methodology [11], and were generated by minimizing several perceptually driven distance metrics [14]. The data-derived gain functions were cataloged in look-up tables indexed by the a posteriori and a priori SNRs, used for augmenting speech in the training acoustic environments. In vehicle contexts, the data-driven method [10] surpassed traditional algorithms (e.g., MMSE) for voice distortion and noise attenuation. The data-driven approach presented in [8] demonstrated superior performance compared to existing state-of-the-art noise reduction algorithms.

The aforementioned data-driven and/or environment-optimized algorithms shown efficacy in enhancing voice quality; nevertheless, their impact on speech intelligibility remains unassessed.

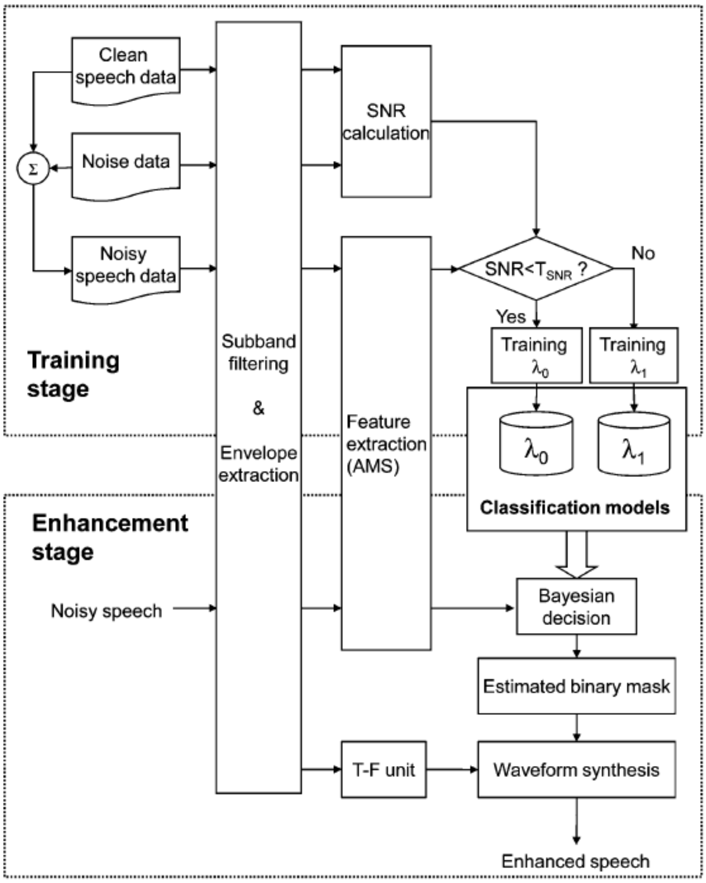

Fig. 1. Block diagram of the training and enhancement stages for the speech

enhancement based on the binary masking of T-F units.

Based on our experience with MMSE-based speech enhancement algorithms [2], we do not anticipate substantial improvements in intelligibility with these methods.

This work adopts an alternative methodology that eschews the development of spectral weighting (gain) functions, instead concentrating on the accurate categorization of spectral SNR in two distinct areas. The adopted methodology is driven by intelligibility research on speech synthesized using the ideal binary mask (IdBM) [15]–[17], which requires access to the signal-to-noise ratio (SNR) at each frequency bin. The ideal binary mask, formerly referred to as the a priori mask, is a method investigated in computational auditory scene analysis (CASA) that preserves the time-frequency regions of the target signal that exhibit a higher signal-to-noise ratio (SNR dB) than the interfering noise, while eliminating the regions that demonstrate a lower SNR dB than the interfering noise. Prior research has shown that the multiplication of the ideal binary mask with the noise-masked signal may provide significant improvements in intelligibility, even at very low signal-to-noise ratio levels of 5 to 10 dB [15], [16]. These investigations required previous knowledge of the actual spectral signal-to-noise ratio and, therefore, the optimal binary mask. In reality, the binary mask must be derived from the corrupted signal, necessitating a precise estimation and classification of the spectral signal-to-noise ratio (SNR). In our prior research [19], we introduced a voice enhancement technique that calculates the binary mask using a Bayesian classifier and synthesizes the improved signal by binary masking (i.e., multiplying the noisy spectra by a binary gain function). This technique disaggregates the input signal into time-frequency units with a rudimentary auditory-like filter bank and employs a basic binary Bayesian classifier to preserve target-dominant time-frequency units while eliminating masker-dominant units. Amplitude modulation spectrograms (AMS) were used as features for training Gaussian mixture models (GMMs) to function as classifiers. In contrast to the majority of speech enhancement methods, the suggested technique does not need speech/noise identification or the estimate of noise statistics. This approach was assessed using listening tests and shown significant improvements in speech intelligibility at very low signal-to-noise ratio levels. The listening tests concentrated on very low signal-to-noise ratio (SNR) levels (e.g., 5 dB), akin to those seen in military contexts, dining establishments, and industrial environments, since speech intelligibility for those with normal hearing is recognized to deteriorate predominantly at these low SNR levels.

BINARY-MASK BASED SPEECH ENHANCEMENT ALGORITHM

Figure 1 illustrates the block structure of the proposed algorithm [19], including a training phase (upper section) and an intelligibility improvement phase (lower section). During the training phase, features are taken from a substantial speech corpus and then used to train two Gaussian mixture models (GMMs) that represent two feature classes: target speech predominating over the masker and the masker predominating over the target speech. In [21], harmonicity-based characteristics were directly retrieved from the voice stream and used in a Bayesian classifier to predict the binary mask. The reliability of harmonicity cues is mostly contingent upon the pitch estimation technique, which often exhibits inaccuracy in low signal-to-noise ratio (SNR) situations.

This work utilizes AMS as characteristics due to their neurophysiological and psychoacoustic foundations [20], [22]. During the enhancement phase, a Bayesian classifier categorizes the time-frequency units of the noise-masked signal into two classifications: target-dominated and masker-dominated.

Individual T-F units of the noise-masked signal are preserved if deemed target-dominated and discarded if categorized as masker-dominated, later used to reconstruct the improved speech waveform. Feature Extraction

The noisy speech signal is first subjected to bandpass filtering into 25 channels based on mel-frequency spacing, covering a bandwidth of 6 kHz (68.5–6000 Hz). The sampling frequency was 12 kHz. The envelopes in each subband are calculated by full-wave rectification and then decimated by a factor of three. The fragmented envelopes are divided into overlapping parts of 128 samples (32 ms) with a 50% overlap. Each segment is subjected to a Hann window and then converted using a 256-point fast Fourier transform (FFT) after zero-padding. The FFT calculates the modulation spectrum for each subband, achieving a frequency resolution of 15.6 Hz.

ADAPTATION TO NEW NOISE ENVIRONMENTS

In the preceding section, we detailed the improvement of noise-masked speech by the estimation of binary masks.

Despite the commendable performance achieved with GMMs trained in various listening situations [19], a user may face a novel sort of noise that is absent from the multiple-noise training set. There are several methods for managing a novel loud environment.

One method involves using a multi-style noise model trained on various noise kinds. We attempted this strategy, but the performance was inadequate. An alternate method involves adjusting the model parameters to suit the new environment.

To swiftly adapt to a new noisy environment, we propose progressively modifying the GMM parameters to integrate the new data, beginning with an initial model trained on limited data. Five Subsequently, we delineate the incremental GMM adaption methodology used. In contrast to the batch-training method, which requires access to the whole dataset, the incremental training strategy perpetually adjusts the model parameters as fresh data becomes available. Thus, the computational burden of the incremental method is less than that of the batch-training method.

A. Preliminary Model

It is assumed that a limited quantity of speech data captured in a calm environment is available for the training of the preliminary model. This data may be retained in memory. In a novel auditory setting, noise-only data are gathered and combined with 10 phrases of clear speech (retained in memory) at signal-to-noise ratio levels of 5, 0, and -5 dB. The distribution of each class may be represented using a limited number of mixture components (e.g., 8), considering the modest quantity of training phrases (e.g., 10 sentences). While the technique of partitioning or augmenting Gaussian mixtures may be used to enhance the number of mixture components as more data is acquired, we opted for a more straightforward approach by training the GMMs with 256 mixture components from the outset.

In the selected incremental training strategy, we just update the parameters of each Gaussian while maintaining a fixed number of mixes. The preliminary model was constructed over the following two phases. Initially, 32 distinct eight-mixture models were generated using the same training data by redoing the original training process 32 times. During each training cycle, the starting centroids for the k-means clustering are selected randomly, resulting in 32 distinct models. In the second stage, the preliminary model with 256 mixes is established by consolidating the 32 models trained with eight mixtures each. The same training data used for all eight-mixture models indicates that the first 256-mixture model has considerable redundancy, implying that several Gaussian components are similar. The redundancy of the models is examined and elaborated upon in Section IV-A.

RESULTS

We credit the considerable increases in intelligibility achieved with the proposed approach (Section II) to the correct categorization of T-F units into target-dominant and masker dominant T-F units. To assess the precision of the GMM-based SNR classifier, we calculated the hit rate (HIT) and false alarm rate (FA) using the same test sets used in the listening trials. The classification accuracy, represented by HIT and FA, of the trained GMM classifiers is shown in Table I as a function of the number of accumulated phrases used in the training. We also computed the error rates evaluating the classifier's performance without differentiating between miss and false alarm errors. A significant decrease in error rates, calculated relative to ten-sentence models, was observed across all three tested maskers, ranging from a 34% reduction (achieved with train noise using 200-sentence models) to a 38% reduction (achieved with babble using 200-sentence models).

The detection rates increased with the inclusion of more training data, resulting in an enhanced hit rate and, in most instances, a reduction in the false alarm rate. Perceptually, the two forms of mistakes that may arise, namely miss (-HIT) and false alarm, are not equal [16]. This occurs because false alarm mistakes may create additional noise distortion, since time-frequency units that would typically be nullified (probably associated with the masker) would now be preserved. The omission errors will likely result in voice distortion, since these mistakes cause the elimination of time-frequency units dominated by the target signal, which should be preserved. To address the cumulative impact of both errors (misses and false alarms), we recommend employing the difference metric, HIT-FA. Table I presents the difference measure as a function of the amount of aggregated phrases used in the training. The difference measure value grows with the inclusion of more training data, indicating a potential association with speech intelligibility ratings. We calculated the connection between the difference metric and speech intelligibility scores.

CONCLUSION

Significant improvements in intelligibility were attained using the suggested approach utilizing a restricted amount of training data. Under typical circumstances, a minimum of 80 sentences was shown to be enough for achieving significant improvements in intelligibility. The clarity of voice processed by the suggested algorithm was much greater than that attained by human listeners perceiving raw (corrupted) speech. This is due to the precise categorization of T-F units into target-dominated and masker-dominated categories, followed by a dependable estimate of the binary mask. The precise categorization of time-frequency (T-F) units into target and masker-dominated units was achieved by the use of neurophysiologically inspired features (AMS) and meticulously crafted Bayesian classifiers (GMMs). In contrast to the mel-frequency cepstrum coefficients (MFCCs) [35] often used in speech recognition, the amplitude modulation spectrum (AMS) characteristics include

data about amplitude and frequency modulations, which are recognized as essential for speech intelligibility [36]. A quantifiable metric derived from classification accuracy (HIT-FA)
Intelligibility of speech generated by algorithms that compute the binary mask. This metric was determined to consistently predict speech intelligibility. The study's results indicate that algorithms capable of consistently estimating or classifying the SNR in each time-frequency unit may enhance speech intelligibility.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC, 2007.

[2] Y. Hu and P. C. Loizou, “A comparative intelligibility study of singlemicrophone noise reduction algorithms,” *J. Acoust. Soc. Amer.*, vol. 122, pp. 1777–1786, 2007.

[3] J. S. Lim, “Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-26, no. 5, pp. 471–472, Oct. 1978.

[4] J. A. Zakis, H. Dillon, and H. J. McDermott, “The design and evaluation of a hearing aid with trainable amplification parameters,” *EarHear.*, vol. 28, no. 6, pp. 812–830, 2007.

[5] J. E. Porter and S. F. Boll, “Optimal estimators for spectral restoration of noisy speech,” in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1984, pp. 18A.2.1–18A.2.4.

[6] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[7] J. Erkelens, J. Jensen, and R. Heusdens, “A general optimization procedure for spectral speech enhancement methods,” in *Proc. Eur. Signal Proc. Conf.*, Florence, Italy, Sep. 2006.

[8] J. Erkelens, J. Jensen, and R. Heusdens, “A data-driven approach to optimizing spectral speech enhancement methods for various error criteria,” *Speech Commun.*, vol. 49, pp. 530–541, 2007.

[9] T. Fingscheidt and S. Suhadi, “Data-driven speech enhancement,” in *Proc. ITG-Fachtagung Sprachkommunikation*, Kiel, Germany, 2006.

[10] T. Fingscheidt, S. Suhadi, and S. Stan, “Environment-optimized speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[11] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[12] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model,” *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[13] C. Bin and P. C. Loizou, “A Laplacian-based MMSE estimator for speech enhancement,” *Speech Commun.*, pp. 134–143, 2007.

[14] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[15] D. Brungart, P. Chang, B. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

[16] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary- masked speech: Implications for noise reduction,” *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[17] N. Li and P. C. Loizou, “Effect of spectral resolution on the intelligibility of ideal binary masked speech,” *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 59–64, 2008.

[18] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[20] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1593–1602, 1994.