

RESEARCH

Open Access

A perceptual masking approach for noise robust speech recognition

Hari Krishna Maganti* and Marco Matassoni

Abstract

This article describes a modified technique for enhancing noisy speech to improve automatic speech recognition (ASR) performance. The proposed approach improves the widely used spectral subtraction which inherently suffers from the associated musical noise effects. Through a psychoacoustic masking and critical band variance normalization technique, the artifacts produced by spectral subtraction are minimized for improving the ASR accuracy. The popular advanced ETSI-2 front end is tested for comparison purposes. The performed speech recognition evaluations on the noisy standard AURORA-2 tasks show enhanced performance for all noise conditions.

1 Introduction

Enhancement of noise corrupted speech signals is a challenging task for speech processing systems to be deployed in real-world applications. In practice, speech signals are usually degraded by additive background noises, reverberation effects and speech signals from other speakers [1]. The primary goal of robust speech processing techniques is to improve intelligibility and quality of noise corrupted speech in perspective human listeners and modify the same and extract robust features that lead to improved performance for speech recognition systems.

Apart from extracting robust features which represent parameters less sensitive to noise by modifying the extracted features [1], other research directions aimed at increasing the performance of speech recognizers in noise are: speech signal enhancement, model adaptation and hybrid methods [2-5]. The model adaptation techniques fail to perform in constantly changing environments where little or no adaptation data is available and hybrid methods attempt to preprocess speech signals and depend on reliability of estimations of those segments. On the other hand the signal enhancement techniques require no training, and provide “real-time” enhancement of the recognition accuracy. The spectral subtraction method of speech enhancement is the most widely used conventional method for reducing additive noise. Many improvements are proposed to deal with the problems typically associated to spectral subtraction such as residual broadband

noise and narrow band tonal noise referred as musical noise [4]. Other variants of spectral subtraction for this purpose include spectral subtraction with over subtraction, non-linear spectral subtraction, multi-band spectral subtraction, MMSE spectral subtraction, and extended spectral subtraction [3,6].

Spectral subtraction based on perceptual properties has been investigated to improve intelligibility and quality of the speech signals [7-9]. The masking properties of human auditory system are incorporated into the enhancement process in order to attenuate the noise components that are already inaudible due to masking. In [7], the selected masking threshold level is high, so that the residual noise will be masked and will be inaudible. In [8], a psychoacoustical spectral weighting rule is proposed which utilizes only estimates of the masking threshold and noise power spectral density for complete masking of distortions of the residual noise. The application of using human auditory masking in Kalman filtering to speech enhancement is considered in [9]. Furthermore, another novel approach based on sub-band variance normalization technique was proposed where speech frames are characterized by high variance and noise frames by low variance, which are suppressed to improve the ASR performance in presence of both additive noise and reverberation [10].

In the present study, an alternate approach based on psycho-acoustical model for reducing the artifacts associated with spectral subtraction for improving speech recognition performance in the presence of additive noise is proposed. Based on the human auditory system, the

*Correspondence: maganti@fbk.eu
Fondazione Bruno Kessler CIT-irst, via Sommarive 18, 38123 Trento, Italy

noise below the audible threshold is suppressed which reduces the amount of modification to the spectral magnitude, and hence the amount of distortion introduced into the cleaned speech signal. Further, critical band variance normalization is performed to minimize the musical noise which is caused by increased variance at random frequencies. The features derived from the combination of techniques are effective in providing robustness. The studied features are shown to be reliable and robust to the effects of the additive noise. The effectiveness of the proposed features is demonstrated with experiments on noisy AURORA-2 database. For comparison purposes the recognition results obtained by using the standard spectral subtraction and ETSI advanced front-end are tested [11].

The article is organized into the following sections. Section 2 briefly summarizes the principle spectral subtraction using over subtraction. Section 3 describes psycho-acoustic motivated features including tone and noise masking and critical band variance normalization. Section 4 discusses the spectral subtraction with perceptual post-filter and Section 5 describes the database, experiments and results. Finally, Section 6 concludes the article.

2 Spectral subtraction for speech enhancement

Spectral subtraction algorithms are based on subtraction of the noise estimate from original speech. The power spectral density (PSD) of a clean speech signal is estimated by subtracting the PSD of the noise from the PSD of the noisy speech signal [12]. Each estimate of the PSD is performed within a short-time segment based on the assumption that noise is stationary or slowly varying, and that the noise spectrum does not change significantly between the updating periods. The noisy speech signal can be expressed as

$$y(t) = s(t) + d(t) \quad (1)$$

where $y(t)$ is the degraded speech signal, $s(t)$ represents the clean signal, $d(t)$ is the additive noise, which is uncorrelated with the speech signal and unknown.

The power spectrum of the noisy speech signal is represented as:

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 + S(\omega) \cdot D(\omega)^* + S(\omega)^* \cdot D(\omega) \quad (2)$$

where $D(\omega)^*$ and $S(\omega)^*$ represent the complex conjugates of $D(\omega)$ and $S(\omega)$, respectively. If noise and speech are uncorrelated then the terms $S(\omega) \cdot D(\omega)^*$, $S(\omega)^* \cdot D(\omega)$ can be neglected. The power spectral subtraction estimate of the enhanced speech $S(\omega)$ is given by

$$|\hat{S}(\omega)|^2 \approx |Y(\omega)|^2 - |\bar{D}(\omega)|^2 \quad (3)$$

where $|\bar{D}(\omega)|^2$ is the average value of noise square-magnitude taken during non-speech activity. The performance of this technique depends on the accuracy of noise estimation and is limited by the processing distortions caused by random variations of the noise spectrum. The non-linear mapping of spectral estimates that fall below a threshold, where noise has been overestimated results in some randomly located negative values for the estimated clean speech magnitude. This leads to undesired residual noise called musical noise (narrow band spectrum with randomly distributed tones over time and frequency).

An important variation of spectral subtraction proposed in [4] in order to alleviate the problem of musical noise is

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\bar{D}(\omega)|^2, & \text{if } |\hat{S}(\omega)|^2 > \beta|\bar{D}(\omega)|^2 \\ \beta|\bar{D}(\omega)|^2, & \text{otherwise} \end{cases} \quad (4)$$

where α , β are the subtraction factor and spectral floor parameter. To reduce the speech distortion caused by large values of α , its value is adapted from frame to frame depending on the segmental noisy signal to noise ratio (NSNR) of the frame. In general, less subtraction is applied for frames with high NSNR and vice versa. For the estimate of the noise power spectrum, minimum statistics technique is used [13]. The NSNR is computed using a decision directed approach as proposed in [14].

As described in [4], the oversubtraction factor α can be calculated as

$$\alpha_m = \begin{cases} 1, & \text{NSNR}_m \geq 20 \text{ dB} \\ \alpha^0 - \frac{3}{20} \text{NSNR}_m, & -6 \text{ dB} \leq \text{NSNR}_m < 20 \text{ dB} \\ 4.9, & \text{NSNR}_m < -6 \text{ dB}, \end{cases} \quad (5)$$

where m is the frame index and α^0 is the value of α at 0 dB NSNR. With the higher over-subtraction, the stronger components with a low SNR are attenuated which prevent musical noise. But, too strong over-subtraction may suppress too many components causing distortion to the signal.

3 Psychoacoustical masking model

The oversubtracted spectral subtraction reduces the noise to some extent but the musical noise is not completely eliminated, effecting the quality of the speech signal. There is a trade-off between the amount of noise reduction and speech distortion. The perceptual based techniques help in reducing the noise by taking advantage of the masking properties of the auditory system. In order to further enhance the quality, the noise and tones are masked and critical band variance normalization is performed by incorporating the masking properties of human

auditory system. The human auditory system does not perceive all the frequencies in the similar way, and is limited to mask certain sounds in the presence of competitive sounds. The two main properties of the human auditory system that make up the psychoacoustic model are: absolute threshold of hearing and auditory masking.

3.1 Absolute threshold of hearing

The absolute threshold of hearing defines the minimum sound level that can be heard by an active listener in absence of other sounds. The quiet threshold can be expressed by the following equation:

$$ATH(f) = 3.64 \left(\frac{f}{1000} \right)^{-0.8} - 6.5e^{-0.6 \left(\frac{f}{1000} - 3.3 \right)^2} + 10^{-3} \times \left(\frac{f}{1000} \right)^4 \quad (6)$$

The frequency components of the signal with power levels that fall below the absolute threshold of hearing (ATH) can be discarded, as they do not contribute in improving perceptibility of the signal.

3.2 Auditory masking

The inability of the human auditory perception system to identify the minute differences in frequency when played at the same time is known as masking. A strong 1 kHz signal masks the nearby frequencies, making them inaudible to the listener. For a masked signal to be heard, its power has to be increased to a level greater than that of a threshold that is determined by the frequency of the masker tone and its strength.

The masking analysis method described in MPEG1 audio coder is used to detect the tonal and nontonal components [15]. The tonal and noise masking threshold that

give the maximum level of noise that is inaudible in the presence of speech is computed. The calculation steps as described in [15] are:

The frequency analysis along the critical band scale is performed by mapping the power spectrum onto an auditory frequency axis, by combining FFT bins into equally-spaced intervals on the Bark scale defined by:

$$Bark(f) = 13 \arctan(0.00076f) + 3.5 \arctan \left(\left(\frac{f}{7500} \right)^2 \right) \quad (7)$$

where f is the frequency in linear domain. The relationship between frequency in Hertz and Bark scales is shown in Figure 1.

The masking threshold is computed from the short-term power spectral density estimate of the input signal. The power density spectrum is obtained from the FFT of the input signal, following multiplication by a Hann window. The magnitude of each spectral component is converted to a decibel scale, to obtain the estimate $P[k]$. The power spectrum is normalized to a level of 96 dB SPL, such that the maximum spectral component corresponds to this value.

The following step involves the discrimination between tonal and noise maskers. This accounts for the dependence of masking thresholds on the nature of the maskers. First, tonal components are identified through the detection of local maxima within the power spectrum. A component is labeled as a local maximum if $P[k] > P[k-1]$ and $P[k] \geq P[k+1]$. Components are declared as tonal if $P[k] - P[k+j] \geq 7$ dB, where j lies within a neighborhood that is dependent on the center frequency, k . The sound pressure level of the tonal masker, $P_{TM}(z)$, where z is the Bark value of the frequency bin k , is computed as follows:

$$P_{TM}(z) = 10 \log_{10} \left(10^{P[k-1]/10} + 10^{P[k]/10} + 10^{P[k+1]/10} \right) \quad (8)$$

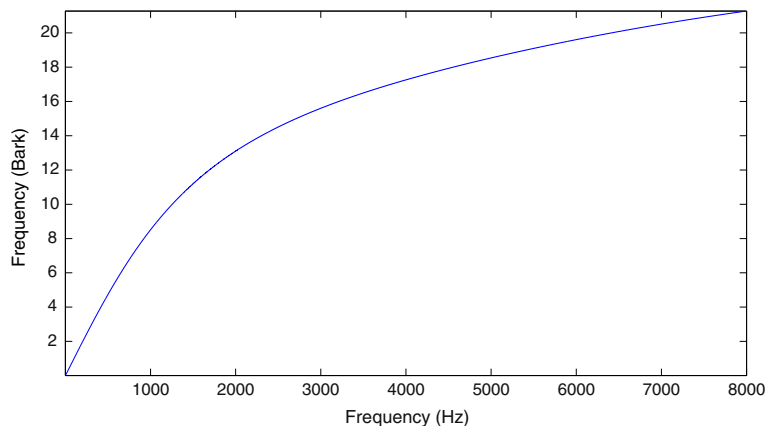


Figure 1 Relationship between frequency in Hertz and Bark scales.

Tonal maskers are removed from the power spectrum $P[k]$, by setting all frequency lines within the examined range to $-\infty$. The sound pressure levels of noise maskers are obtained by summing the energies of spectral lines within each critical band to compute $P_{TM}(z)$.

Subsequently, the number of maskers considered for threshold computation is reduced. At first, only maskers having a level above the absolute threshold of hearing are retained. A decimation process then occurs between multiple tonal maskers that lie within half of a critical band. The tonal masker having the highest level is maintained while the other elements are removed from the directory of maskers. The tone and noise maskers for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR is shown in Figure 2.

The contribution of each masker to the overall masking threshold is evaluated. The spread of masking effects is modeled using the spreading function described below:

$$s(z_j, \Delta z, P_M(z_j)) = \begin{cases} 17\Delta z - 0.4P_M(z_j) + 11, & -3 \leq \Delta z < -1 \\ (0.4P_M(z_j) + 6)\Delta z & -1 \leq \Delta z < 0 \\ -17\Delta z & 0 \leq \Delta z < 1 \\ (0.15P_M(z_j) - 17)\Delta z - 0.15P_M(z_j) & 1 \leq \Delta z < 8 \\ -\infty & \text{otherwise,} \end{cases} \quad (9)$$

where z_j , Δz , and $P_M(z_j)$ represent, respectively, the masker Bark frequency, the Bark frequency separation between the masker and target and the sound pressure level of the masker. The spread of masking is only considered within the range of $-3 \leq \Delta z < 8$, for reasons of implementation complexity.

The masking indices (in dB) for tonal maskers $a_{TM}(z)$, and noise maskers $a_{NM}(z)$, expressed below in Equations (10) and (11) are both frequency dependent. They represent the offset to be subtracted from the excitation pattern of the masker to obtain the masking pattern. The checked

tone and noise maskers for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR is shown in Figure 3.

$$a_{TM}(z) = -1.525 - 0.257z - 4.5 \quad (10)$$

$$a_{NM}(z) = -1.525 - 0.175z - 0.5 \quad (11)$$

The individual masking thresholds from each masker calculated according to:

$$M_{TM}(z_j, \Delta z) = P_{TM}(z_j) + a_{TM}(z_j) + s(z_j, \Delta z, P_{TM}(z_j)) \quad (12)$$

$$M_{NM}(z_j, \Delta z) = P_{NM}(z_j) + a_{NM}(z_j) + s(z_j, \Delta z, P_{NM}(z_j)) \quad (13)$$

The individual making thresholds are computed for each sub-band, using all the maskers. The global masking threshold per sub-band is computed by summing the individual masking contributions from each masker along with the absolute threshold of hearing, $T_Q(z)$

$$M_g(z_i) = 10 \log_{10} \left(10^{T_Q(z_i)/10} + \sum_{j=1}^m 10^{M_{TM}(z_j, z_i)/10} + \sum_{j=1}^n 10^{M_{NM}(z_j, z_i)/10} \right) \quad (14)$$

The ATH, original PSD, noise and tone maskers, and global masking threshold for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR is shown in Figure 4.

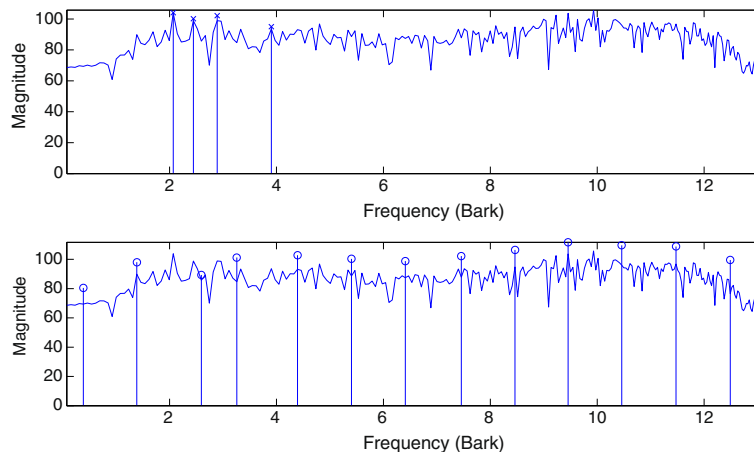


Figure 2 Tone and noise maskers for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR.

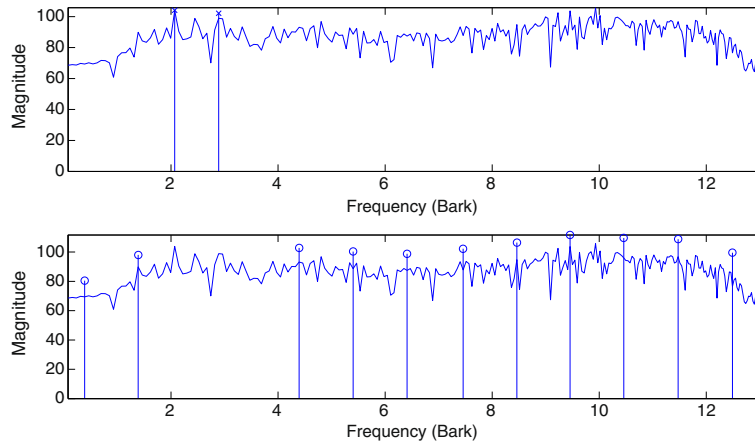


Figure 3 Checked tone and noise maskers for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR.

3.3 Variance normalization

Further, to reduce the effect of any present tones which are caused by increased variance at random frequencies, variance normalization across the critical bands is performed [10]. The variance is computed as

$$v(m) = \frac{1}{K-1} \sum_{i=1}^K (v_i(m) - \hat{v}(m))^2 \quad (15)$$

where K is the number of bands, m is the frame index, \hat{v} is the mean, and v_i is the element number i . To suppress the peaks of noise, these values are normalized with respect to the maximum value across the bands

$$w(m) = \frac{v(m)}{\max\{v(m)\}} \quad (16)$$

4 Spectral subtraction with perceptual post-filter

The description provided in previous Section 3 provided a theoretical foundation for combining the masking properties of human auditory system in a spectral subtraction framework used for speech recognition. The proposed feature extraction methodology is shown in Figure 5. First, the 8 kHz speech signal undergoes pre-emphasis and short segments of speech are extracted with a 25 ms rectangular window, shifted by 10 ms. After the spectral subtraction (spectral floor parameter $\beta = 0.02$, α as defined in Equation (5) with $\alpha^0 = 4$) with oversubtraction, the signal is then filtered by a bank of 24 critical-band bark filters. The global masking threshold is applied, then variance normalization is performed to further suppress the tones at random frequencies. Later, these normalized values are used as weights which are then multiplied with the filter bank energies as shown

$$\tilde{Y}_k(m) = Y_k(m) \cdot w(m) \quad (17)$$

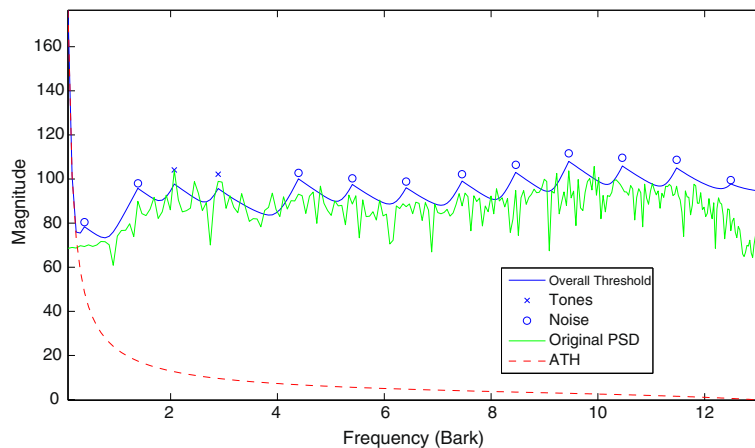
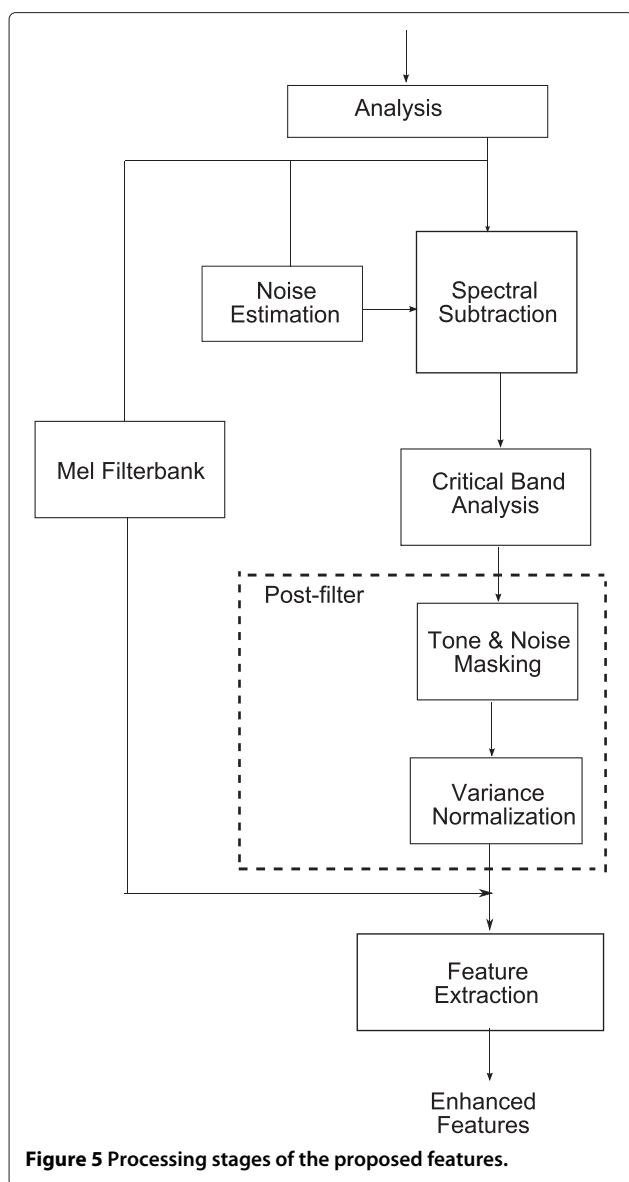


Figure 4 The ATH, original PSD, noise and tones, and global masking threshold for part of the utterance “three six seven” corrupted with subway noise of 5 dB SNR.



MFCC features are extracted from the log Mel spectrum by applying a discrete cosine transformation. First thirteen cepstral coefficients along with their first- and second-order derivatives are used.

5 Experiments and results

The ASR experiments are performed with the proposed approach using a full HTK based recognition system [16] on connected digit recognition task using the Aurora 2 database [17]. The database was designed to evaluate the front-end of ASR systems in noisy conditions, and the training and testing follow the specifications described in [17]. The task is speaker independent connected digit recognition.

Testing data include eight types of realistic background noise subway, babble, car, exhibition hall, restaurant, street, airport and train station noise at various SNRs (clean, 20, 15, 10, 5, 0, and -5 dB). There are three test sets. Set A contains 4004 utterances in the first four types of noise, set B contains 4004 utterances in the other four, and set C contains 2002 utterances, where only subway and street noise are present.

The Aurora 2 task defines two different training modes: training on clean condition only, and training on multi-condition which include both clean and noisy conditions. Experiments with training on both conditions are considered to demonstrate the efficiency of the proposed approach. A word-based ASR system for digit string recognition where each HMM word model has 16 emitting states is adopted. A three-state silence model and a one state short pause model are used. The details about the training set, test set and HTK recognizer can be found in [17]. Results on Aurora 2 test sets A, B, and C with training on clean data and multi-condition data are reported in Figures 6 and 7, respectively.

From Figures 6 and 7, it can be seen that spectral subtraction has the highest word error rates (WER) for both clean and multi-condition training. Also, it can be observed that for a minimal average loss in case of clean speech for both clean and multi-condition training, an improvement is obtained with the proposed features for all the noise conditions. The improvement is particularly large for -5 , 0, 5, 10 and 15 dB SNRs. It can also be observed that the improvements are better for training on clean data than multi-condition data which is consistent with [17]. For both the cases, the approach is precisely able to remove noise as much as possible improving the recognition accuracy.

Figure 8 shows comparison of the performance for average calculated over 0 to 20 dB SNR's for all the test sets A, B, and C, respectively. The first and second bars indicates the average result obtained with spectral subtraction and ETSI-2 frontend trained on clean data. The third bar denotes the performance of proposed approach for clean condition training. The fourth, fifth, and sixth bars show the results obtained with spectral subtraction, ETSI-2 frontend and proposed approach for multi-condition training. It can be observed that for all the test sets the performance of the proposed approach is better than spectral subtraction and ETSI-2 frontend for both training on clean data and multi-condition data.

It can be clearly observed from the Figures 6, 7, and 8 that the performance of proposed features is consistently the best for all noise conditions irrespective of the training. To overcome the basic limitation and application of spectral subtraction technique for recognition tasks, the combination of psychoacoustical masking and critical band variance normalization is effective in minimizing the

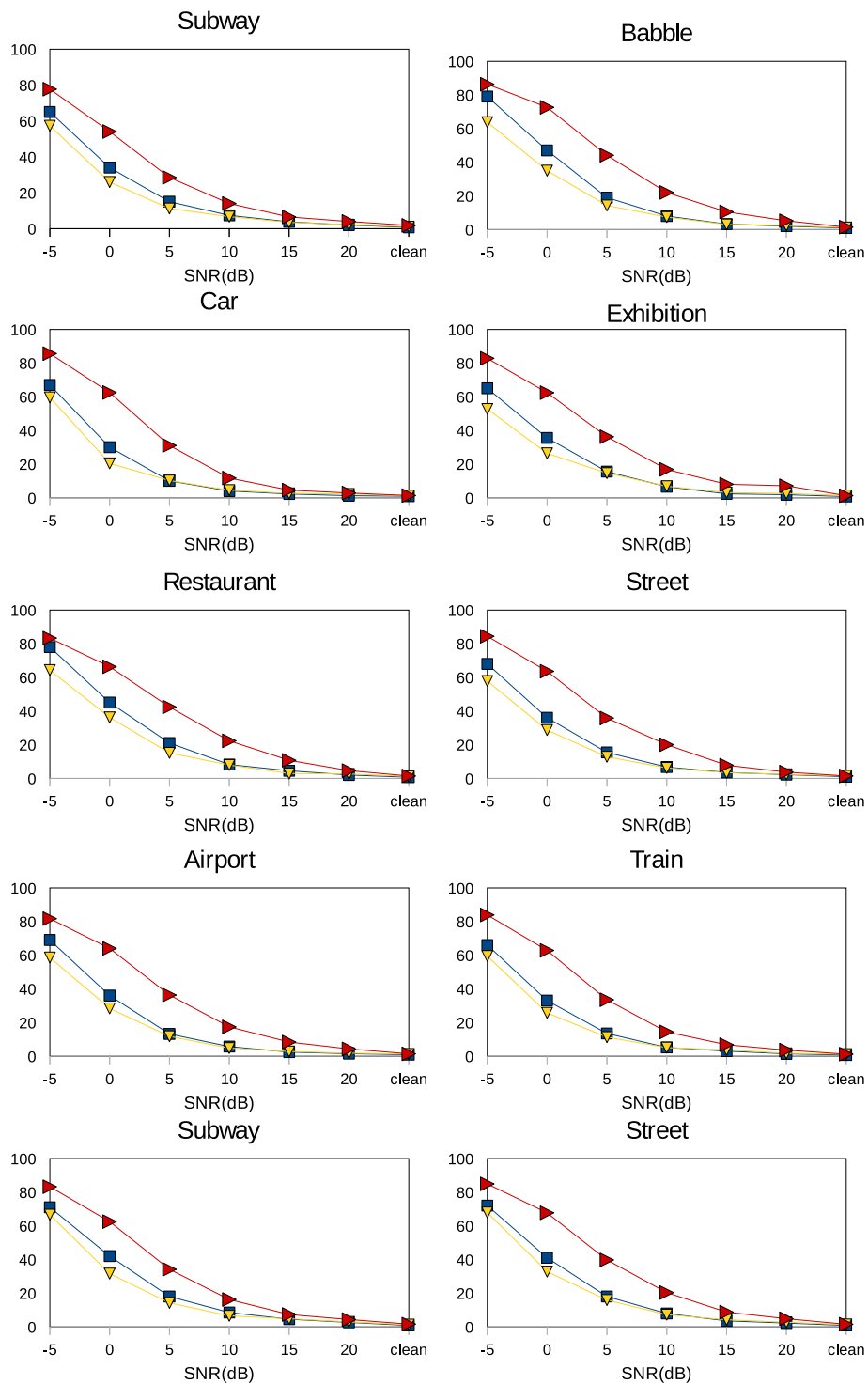


Figure 6 Comparison of WER (%) performance for ETSI-2 (□), proposed features (▽) and spectral subtraction (▷) for all noises of test sets A, B, and C trained on clean data.

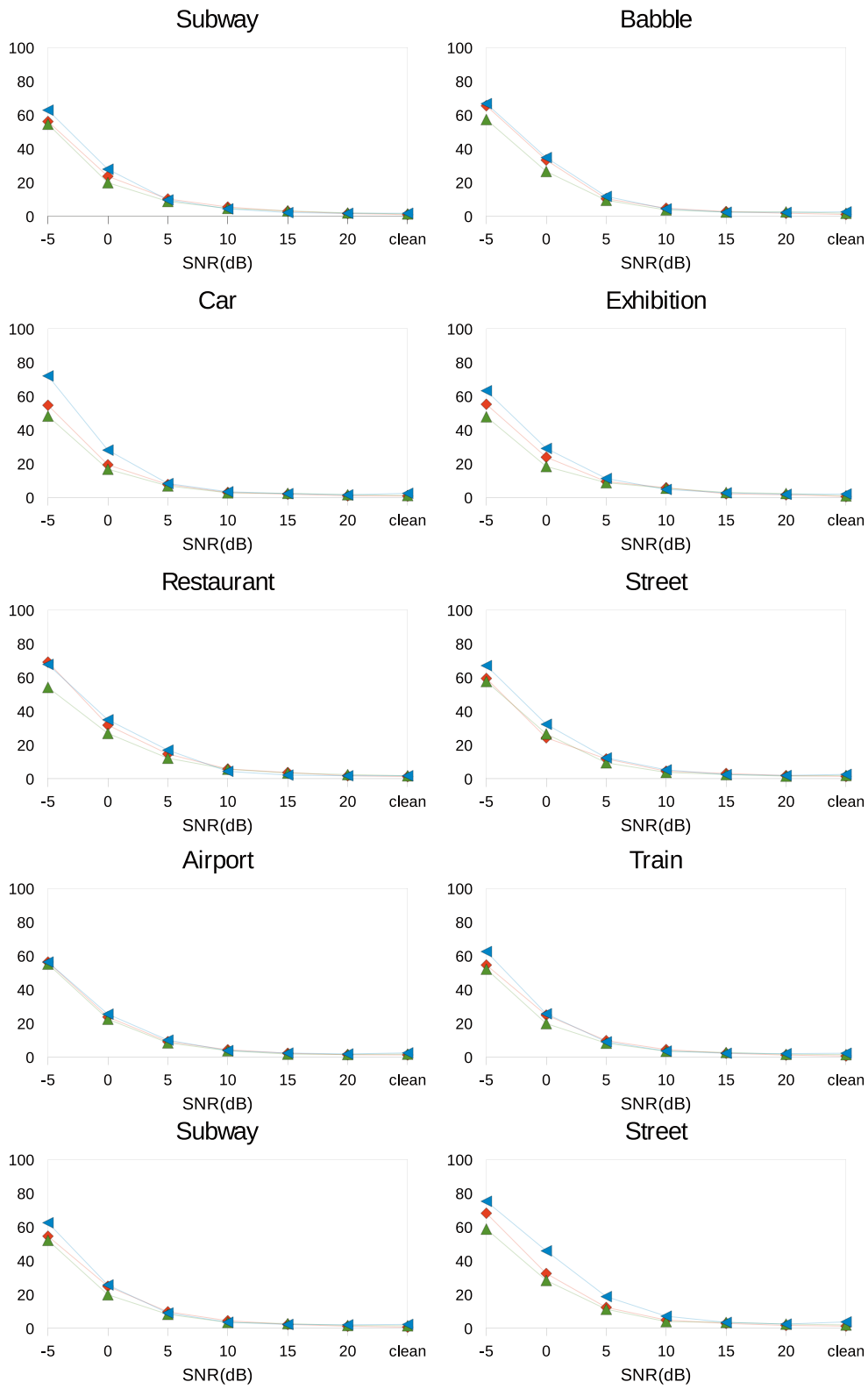


Figure 7 Comparison of WER (%) performance for ETSI-2 (◇), proposed features (Δ) and spectral subtraction (◄) for all noises of test sets A, B, and C trained on multi-condition data.

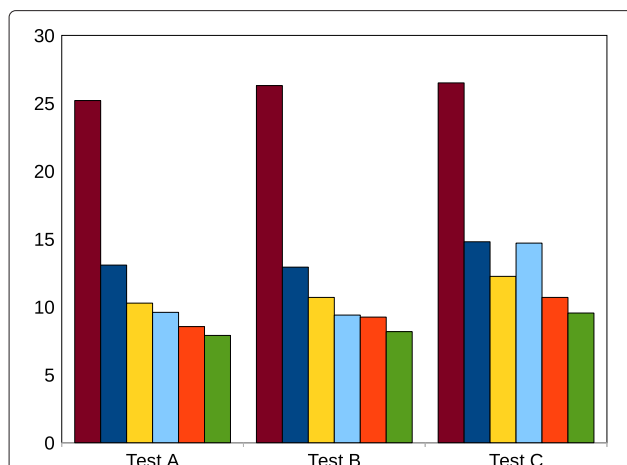


Figure 8 Comparison of average WER (%) performance calculated from 0 to 20 dB for spectral subtraction, ETSI-2 frontend and proposed approach for test sets A, B, and C. The first and fourth and second and fifth bars indicate performance of spectral subtraction and ETSI-2 frontend for training on clean data and multi-condition data. The third and sixth indicate performance of proposed approach for training on clean data and multi-condition data.

artifacts without causing distortion to the original speech signal, thereby improving the ASR accuracy.

6 Conclusions

This article presented a psychoacoustical masking and critical band variance normalization based spectral subtraction approach to improve the speech recognition performance in noisy environments.

The spectral subtraction method was used to reduce the broadband noise due to peaks, and the combination of masking and variance normalization technique was effective in reducing the artifacts by reducing the dynamic range of its magnitude spectrum, which resulted in the improved speech recognition performance. The proposed approach has been evaluated on a standard Aurora-2 database. Results were compared with standard ETSI-2 advanced front-end which show that the proposed features perform consistently better both in terms of robustness and reliability for all types of noises.

In future investigations, improvement of auditory based features to deal with both additive noise and reverberation simultaneously will be investigated. Also, evaluation of these features on large vocabulary tasks to deal with real world noisy speech will be studied.

Competing interests

The authors declare that they have no competing interests.

Received: 8 December 2011 Accepted: 12 November 2012
Published: 22 December 2012

References

1. J Droppo, A Acero, in *Springer Handbook of Speech Processing*, ed. by J Benesty, MM Sondhi, and Y Huang (eds). Environmental robustness, chapter 33, (2008), pp. 653–680

2. S Furui, M Sondhi, in *Advances in speech signal processing*. Marcel Dekker, Inc., (New York, p. 1991)
3. Y Ephraim, I Cohen, *Recent advancements in speech enhancement*. (The Electrical Engineering Handbook, CRC Press, 2006)
4. M Berouti, R Schwartz, J Makhoul, in *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. Enhancement of speech corrupted by acoustic noise, (Washington, USA, 1979), pp. 208–211
5. MJF Gales, in *Springer series of Signal, Image and Speech Processing*, vol. XIV, ed. by D Kolossa, RH Umbach (eds). Robust Speech Recognition of Uncertain or Missing Data, chapter 1, (2011), pp. 101–125
6. P Krishnamoorthy, SRM Prasanna, Reverberant speech enhancement by temporal and spectral processing. *IEEE Transactions on Speech Audio Processing*. **17**(2), 253–266 (2009)
7. N Virag, Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech Audio Processing*. **7**(2), 126–137 (1999)
8. S Gustafsson, P Jax, P Vary, in *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics, (Washington, USA, 1998), pp. 397–400
9. N Ma, M Bouchard, RA Goubran, Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations. *IEEE Transactions on Speech Audio Processing*. **14**(1), 19–32 (2006)
10. HK Maganti, S Zanon, M Matassoni, in *Proc. 19th European Signal Processing Conference (EUSIPCO)*. Sub-band spectral variance feature for noise robust ASR, (Barcelona, Spain, 2011), pp. 2114–2118
11. STQ ETSI ES 202050, Distributed Speech Recognition, Advanced Front-End Feature Extraction Algorithm, Compression Algorithm, ETSI ES 202 050 v1.1.3 (2003-11) (2003)
12. S Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech Signal Processing*. **27**(2), 113–120 (1979)
13. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech Audio Processing*. **9**, 504–512 (2001)
14. Y Ephraim, D Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech Signal Processing*. **ASSP-32**(6), 1109–1121 (1984)
15. T Painter, A Spanias, Perceptual coding of digital audio. *Proc. of the IEEE*. **88**, 451–513 (2000)
16. SJ Young, G Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, PC Woodland, *The HTK book (version 3.4)*. (Cambridge University Engineering Department, Cambridge, UK, 2006)
17. HG Hirsch, D Pearce, Applying the Advanced ETSI Frontend to the Aurora-2 task, in version 1.1 (2006)

doi:10.1186/1687-4722-2012-29

Cite this article as: Maganti and Matassoni: A perceptual masking approach for noise robust speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:29.

Submit your manuscript to a SpringerOpen journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com