Interspeech 2018
2-6 September 2018, Hyderabad

Weighting Time-Frequency Representation of Speech using Auditory Saliency for Automatic Speech Recognition

Cong-Thanh Do[†]*, Yannis Stylianou*

Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

Abstract

This paper proposes a new method for weighting two-dimensional (2D) time-frequency (T-F) representation of speech using auditory saliency for noise-robust automatic speech recognition (ASR). Auditory saliency is estimated via 2D auditory saliency maps which model the mechanism for allocating human auditory attention. These maps are used to weight T-F representation of speech, namely the 2D magnitude spectrum or spectrogram, prior to features extraction for ASR. Experiments on Aurora-4 corpus demonstrate the effectiveness of the proposed method for noise-robust ASR. In multi-stream ASR, relative word error rate (WER) reduction of up to 5.3% and 4.0% are observed when comparing the multi-stream system using the proposed method with the baseline single-stream system not using T-F representation weighting and that using conventional spectral masking noise-robust technique, respectively. Combining the multi-stream system using the proposed method and the single-stream system using the conventional spectral masking technique reduces further the WER.

Index Terms: Time-frequency representation, auditory saliency, multi-stream automatic speech recognition, noise robustness, spectral masking

1. Introduction

When there are environmental noises, word error rate (WER) of automatic speech recognition (ASR) system often increases. Various noise-robust ASR techniques were developed to improve ASR system noise robustness. A profound overview of noise-robust techniques for ASR is presented in a recent literature survey [1]. In [1], noise-robust ASR techniques were classified into 5 categories: (i) feature-domain vs. model-domain techniques, (ii) techniques that exploit prior knowledge about the signal distortion, (iii) techniques that incorporate an explicit distortion model to predict the distorted speech from a clean one, (iv) techniques using uncertainty in either model space or feature space, and (v) techniques that use joint model training in which environmental variability in the training data is removed in order to generate canonical models.

Various noise-robust ASR techniques were developed in the feature-domain of ASR system based on knowledge from human auditory system, for instance the use of non-linear frequency axis [2, 3, 4], the application of the principle of temporal processing in frequency bands of speech signal [5, 6, 7], or the use of spectro-temporal features [8, 9, 10]. A recent overview of perceptually-motivated techniques for noise-robust ASR is presented in [11].

Human auditory attention is one of the mechanisms which help human to recognize better speech in noisy and adverse environments. In essence, auditory attention mechanism acts as a selection process that focus both sensory and cognitive resources on the most relevant events in the sound environment [12]. This selection process is a component of sensory attention which provides weighted representation of our environment, biasing perception towards salient events [13].

In this work, we propose a new method for noise-robust ASR based on human auditory attention mechanism. In the proposed method, a time-frequency (T-F) representation of speech is weighted using an auditory saliency map (ASM) [13] which models the mechanism for allocating auditory attention. The ASM is a two-dimensional (2D) matrix which represents the saliency at every location in a T-F representation of speech by a scalar quantity to guide the selection of attended locations, based on the spatial distribution of saliency [13, 14]. Weighting the T-F representation of speech using the ASM is effectively consistent with the fact that attention mechanism provides weighted representation of the environment. We show the effectiveness of the proposed method in improving ASR noise robustness, especially in multi-stream ASR framework [15]. Experiments are performed on Aurora-4 corpus [16] using Kaldi speech recognition toolkit [17].

The paper is organized as follows. Section 2 presents prior work related to the proposed method. In section 3, the algorithm for computing the ASM is presented. The application of auditory saliency in ASR is presented in section 4. Section 5 presents ASR experiments and section 6 concludes the paper.

2. Relation to prior work

Spectral masking is an approach aiming at weighting the T-F representation of noisy speech in order to improve ASR noise robustness [18, 19] and speech intelligibility for humans [20]. In the spectral masking approach, a weight matrix is computed for a T-F representation of speech, for instance the spectrogram or cochleagram. Each element of the 2D T-F representation, or T-F unit, is multiplied with an element in the weight matrix. Values of the elements in the weight matrix are either binary [18, 21] or continuous [22, 23]. The binary values can be considered as a binary approximation of the continuous values. The computation of the weight matrix is often based on energy of clean speech and noise at each T-F unit.

Ideal binary mask (IBM) is a spectral masking technique which uses a binary matrix as weight matrix. This is a typical spectral masking technique for noise-robust ASR [18, 21]. In the binary matrix, a value of 1 denotes that the speech energy $S(n,\omega)$ in the corresponding T-F unit of the T-F representation exceeds the noise energy $N(n,\omega)$ by a predefined threshold θ. Here n represents a time frame and ω represents a frequency band. A value of 0 in the binary matrix denotes otherwise. More specifically, the binary matrix or binary mask $B(n,\omega)$ is defined as

$$B(n,\omega) = \begin{cases} 1 & \text{if } \; S(n,\omega) - N(n,\omega) > \theta \\ 0 & \text{otherwise.} \end{cases}$$

The threshold θ is typically set to 0 corresponding to a local signal-to-noise ratio (SNR) of 0 dB. Speech is expected to be

[†]Corresponding author. E-mail: cong-thanh.do@crl.toshiba.co.uk

10.21437/Interspeech.2018-1721

segregated from noise by applying element-wise multiplication of the binary mask with the T-F representation of noisy speech. In realistic conditions where only noisy speech signal is known, speech and noise signals are separated from the noisy speech signal to estimate the binary mask. The 0s in the estimated binary mask (EBM) are often replaced with an alternative floor value α between 0 and 1 as this was found to improve the overall performance [18, 21]. In this work, the EBM spectral masking technique is used as one of the baselines for comparison with the proposed method based on human auditory attention.

3. Auditory saliency map (ASM)

The ASM is a 2D matrix which represents the saliency at every location in a T-F representation of speech by a scalar quantity to guide the selection of attended locations, based on the spatial distribution of saliency [13, 14]. The ASM extracts in parallel features from speech signal which represent various levels of sound feature analysis by auditory neurons [13]. These features are believed to help human auditory system to detect sounds of interest in a noisy environment [24]. In this respect, the ASM uses different sets of filters to quantify sound intensity, frequency contrast, and temporal contrast, and compares each individual feature across scales using a center-surround mechanism and thresholding [13]. The feature maps created after the comparison are normalized to obtain a feature-independent scale. The ASMs for individual features are created by summing the feature maps at different scales. The overall ASM is created by summing the individual ASMs, in analogy to the idea of feature integration. This model was confirmed to replicate basic properties of auditory scene perception by humans [13]. The algorithm for computing the ASMs from individual features is described in greater detail as follows.

Let $\mathbf{X}(n,\omega)$ denotes the magnitude spectrum or spectrogram of a given speech signal. $\mathbf{X}(n,\omega)$ is a 2D matrix which is computed by applying a discrete Fourier transformation (DFT) to the speech signal on a frame-by-frame basis, and then taking the magnitude of the resulting 2D complex spectrum. The log magnitude spectrum $\widehat{\mathbf{X}}(n,\omega)$ is computed as $\widehat{\mathbf{X}}(n,\omega) = \log(\mathbf{X}(n,\omega))$. $\widehat{\mathbf{X}}(n,\omega)$ is then down-sampled by factors of $2^k, k = 0,...,K-1$ by using polynomial interpolation to create K log magnitude spectra $\widehat{\mathbf{X}}_0(n,\omega), \widehat{\mathbf{X}}_1(n,\omega), ...\widehat{\mathbf{X}}_{K-1}(n,\omega)$. These log magnitude spectra are then filtered by a 2D Gabor filter $\mathbf{G}(n,\omega)$ to create K 2D representations $\mathbf{R}_0(n,\omega), \mathbf{R}_1(n,\omega), ..., \mathbf{R}_{K-1}(n,\omega)$ where $\mathbf{R}_k(n,\omega) = \widehat{\mathbf{X}}_k(n,\omega) * \mathbf{G}(n,\omega), k = 0,...,K-1$. The 2D Gabor filter $\mathbf{G}(n,\omega)$, which is product of a 2D sinusoidal plane wave and a 2D Gaussian envelope [25], is used to approximate the function of the auditory receptive fields [13, 26].

The differences between the representations $\mathbf{R}_k, k = 0,...,K-1$ at different scales are then computed through a center-surround mechanism to mimic the properties of local cortical inhibition [27]. To this end, the 2D representations $\mathbf{R}_k, k = 1,...,K-1$ are first up-sampled by using polynomial interpolation to have the same dimensions as the representation $\mathbf{R}_0$ which has the same dimensions as the original 2D magnitude spectrum $\mathbf{X}(n,\omega)$. Element-wise subtraction is then computed between the up-sampled representations $\widehat{\mathbf{R}}_k, k = 0,...,K-1$ where $\widehat{\mathbf{R}}_0 = \mathbf{R}_0$. More specifically, given a scale $k, k = 0,...,K-3$, the difference is computed between the up-sampled representations at scale k and scales $k+1$ and $k+2$ to create two feature maps $\mathbf{F}_{k,k+1}$ and $\mathbf{F}_{k,k+2}$ as follows: $\mathbf{F}_{k,k+1} = \widehat{\mathbf{R}}_k - \widehat{\mathbf{R}}_{k+1}$ and $\mathbf{F}_{k,k+2} = \widehat{\mathbf{R}}_k - \widehat{\mathbf{R}}_{k+2}$. A threshold is then applied on $2(K-2)$ resulting feature maps to keep only their positive values in the thresholded feature maps $\widehat{\mathbf{F}}_i, i = 0,...,2(K-2)-1$.

The thresholded feature maps $\widehat{\mathbf{F}}_i, i = 0,...,2(K-2)-1$ are then normalized with respect to their local maxima [13, 14]. Given a feature map $\widehat{\mathbf{F}}_i$, the normalized feature map $\widetilde{\mathbf{F}}_i$ is computed as $\widetilde{\mathbf{F}}_i = \frac{\widehat{\mathbf{F}}_i}{\Phi}(1-\bar{\varphi})^2$ where Φ and $\bar{\varphi}$ are two scalar quantities representing the global maximum and the average of the local maxima of the map, respectively. The ASM $\mathbf{S_G}$ for an individual feature, e.g. sound intensity, extracted with the Gabor filter $\mathbf{G}$ is computed as sum of the normalized feature maps: $\mathbf{S_G} = \sum_{i=0}^{2(K-2)-1} \widetilde{\mathbf{F}}_i$. The algorithm for computing the ASM $\mathbf{S_G}$, or individual ASM, is depicted in Fig. 1.

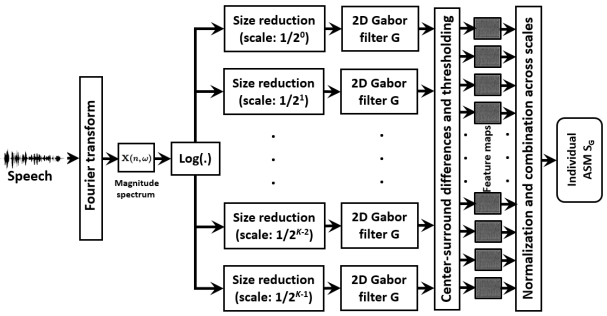

Figure 1: *Algorithm for computing an individual auditory saliency map (ASM)* $\mathbf{S_G}$ *using one Gabor filter* $\mathbf{G}$.

In this paper, three different 2D Gabor filters are used in the computation of three different individual ASMs. These filters are similar to those used in the computation of the ASMs proposed in [13]. Fig. 2(a), 2(b) and 2(c) show these three filters $\mathbf{G_I}, \mathbf{G_F}$ and $\mathbf{G_T}$ which are used to extract features related to sound intensity, frequency contrast and temporal contrast, respectively. The individual ASMs computed by using $\mathbf{G_I}, \mathbf{G_F}$ and $\mathbf{G_T}$ are denoted as $\mathbf{S_{G_I}}, \mathbf{S_{G_F}}$ and $\mathbf{S_{G_T}}$, respectively. The overall ASM $\mathbf{S_O}$ is computed by summing the individual ASMs $\mathbf{S_{G_I}}, \mathbf{S_{G_F}}$ and $\mathbf{S_{G_T}}$ as proposed in [13, 14]:

$$\mathbf{S_O} = \frac{1}{3}(\mathbf{S_{G_I}} + \mathbf{S_{G_F}} + \mathbf{S_{G_T}}).$$

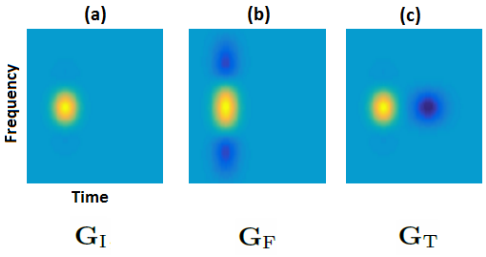

Figure 2: *2D Gabor filters for extracting features related to intensity* ($\mathbf{G_I}$)*, frequency contrast* ($\mathbf{G_F}$)*, and temporal contrast* ($\mathbf{G_T}$) *from the 2D magnitude spectrum of speech.*

Fig. 3 shows examples of the log magnitude spectrum $\widehat{\mathbf{X}}(n,\omega)$ together with the individual and overall ASMs, $\mathbf{S_{G_I}}, \mathbf{S_{G_F}}, \mathbf{S_{G_T}}$ and $\mathbf{S_O}$, computed from a noisy speech signal in the Aurora-4 corpus. The values of the elements in the ASMs are in the range $[0, 1]$. In this work, a 1024-point DFT is used to compute the 2D magnitude spectrum and $K = 6$ is the

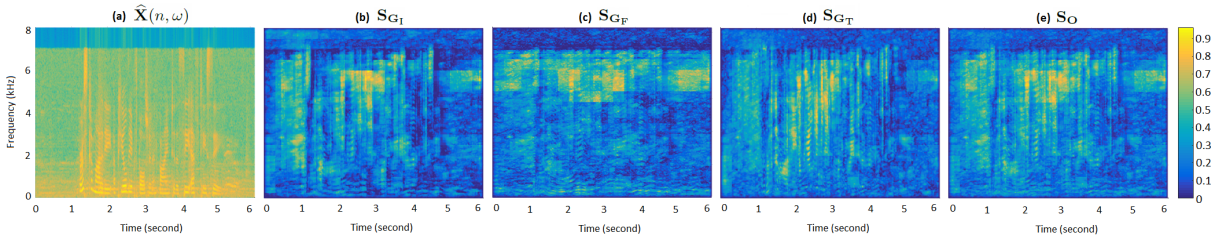

Figure 3: *Examples of the log magnitude spectrum (a) and the individual and overall ASMs (b, c, d, and e) computed from a noisy speech signal in the Aurora-4 corpus.*

number of scales used in the computation of the ASMs as these parameters provide better ASR performance. The ASMs have the same dimensions as the magnitude spectrum.

4. Application of auditory saliency in ASR

The individual and overall ASMs $\mathbf{S}_{\mathbf{G}_\mathrm{I}}, \mathbf{S}_{\mathbf{G}_\mathrm{F}}, \mathbf{S}_{\mathbf{G}_\mathrm{T}}$ and $\mathbf{S}_\mathbf{O}$ are used to weight the 2D magnitude spectrum $\mathbf{X}(n,\omega)$ prior to features extraction for ASR. The weighting is done by element-wise multiplication of each map with the 2D magnitude spectrum. Acoustic features for ASR are then computed from the weighted magnitude spectrum matrix. In this work, Mel filterbank (FBANK) features [28], which are created by skipping the discrete cosine transform (DCT) in the Mel frequency cepstral coefficients (MFCCs) [2] computation, are used. As there is a log function within the FBANK features extraction which is applied on the sums of the element-wise multiplication between the (weighted) magnitude spectrum and Mel-scale filterbank [2], the weighting prior to features extraction should be performed with the linear scale magnitude spectrum, i.e with $\mathbf{X}(n,\omega)$. As a result, an exponential function is applied on the ASMs to transform these maps to linear scale because they are computed from the log magnitude spectrum $\widehat{\mathbf{X}}(n,\omega)$.

4.1. Single-stream ASR

In single-stream hidden Markov model (HMM)-based ASR, FBANK features are computed either from $\mathbf{X}(n,\omega)$, $\mathbf{X}(n,\omega)\odot \exp(\mathbf{S}_{\mathbf{G}_\mathrm{I}})$, $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_{\mathbf{G}_\mathrm{F}})$, $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_{\mathbf{G}_\mathrm{T}})$, or from $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_\mathbf{O})$ where $\odot$ denotes element-wise multiplication. These FBANK features are then used during the training of individual acoustic models and during testing of single-stream ASR systems. In this work, convolutional neural networks (CNNs) [29] are used as acoustic models in hybrid CNN-HMM ASR systems [30, 31].

In addition to the baseline single-stream system in which no weighting is applied, i.e the system uses FBANK features computed from the magnitude spectrum $\mathbf{X}(n,\omega)$, a system using the EBM spectral masking technique is used as another baseline system (see section 2). In this system, the EBM $B(n,\omega)$ is computed from a speech signal, then, it is used to weight the 2D complex spectrum of the speech signal by element-wise multiplication. Subsequently, FBANK features can be computed either from the magnitude of the weighted complex spectrum or from the waveform re-synthesized from the weighted complex spectrum which is a typical way of extracting features when the EBM is used [18, 21]. In this work, FBANK features are extracted from the re-synthesized waveform.

For computing the EBM $B(n,\omega)$, noise-free and noise DFT coefficients are estimated from a noisy speech signal. To this end, noise-free DFT coefficients are estimated using a minimum mean-square error (MMSE) estimator [32] which depends on the noise power spectral density (PSD) estimated by a PSD tracking algorithm proposed in [33]. Noise DFT coefficients are estimated as the difference between the noisy DFT coefficients and the noise-free DFT coefficients due to the linearity of the Fourier transform [33]. Once the noise-free and noise DFT coefficients are estimated, the EBM $B(n,\omega)$ can be computed (see section 2) and used to weight the complex spectrum of noisy speech. The 0s in the binary masks are replaced with an alternative floor value α between 0 and 1 [18, 21]. In the present work, $\alpha = 0.9$ was found to give better ASR performance than other values of α.

4.2. Multi-stream ASR

Multi-stream ASR combines information from different speech recognition streams to improve ASR performance [15]. The combination of different ASR streams can exploit particular strength of each technique, for instance acoustic features, used in each stream. The combination can be performed at features level, output of acoustic models level or lattices level [34]. In this work, the single-stream CNN-HMM ASR systems are combined together in the multi-stream CNN-HMM ASR framework. Each single-stream ASR system uses FBANK features computed either from $\mathbf{X}(n,\omega)$, $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_{\mathbf{G}_\mathrm{I}})$, $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_{\mathbf{G}_\mathrm{F}})$, $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_{\mathbf{G}_\mathrm{T}})$, or from $\mathbf{X}(n,\omega)\odot\exp(\mathbf{S}_\mathbf{O})$ (see section 4.1). The single-stream system applying the EBM technique is also used. The combination is performed either at the output of the CNN acoustic models or at the output of the decoding (see Fig. 4).

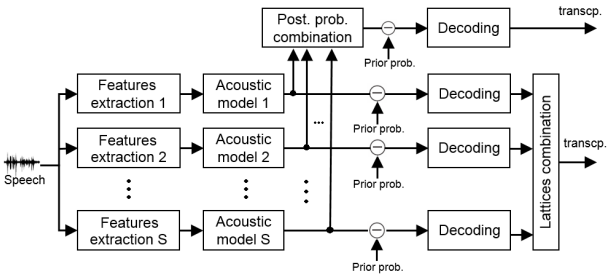

Figure 4: *Multi-stream ASR with combination performed either at the output of the acoustic models which are CNNs or at the output of the decoding. S is the number of streams.*

Posterior probabilities of tied triphone HMM states [31] produced by the CNN acoustic models can be combined by using a number of methods. Here we apply the inverse entropy combination [35] which is one of the effective methods for combining posterior probabilities. In this method, the weight allocated to the posterior probabilities produced by a CNN acoustic model is proportional to the inverse entropy of that acoustic model which characterizes its discriminative capacity [35]. Prior probabilities of tied triphone HMM states are subtracted in log domain from the combined posterior probabilities to get

the scaled log-likelihoods [30] which are subsequently used for decoding. The combination can also be performed on the lattices obtained after the decoding. In this work, lattices are combined based on Bayes risk minimization [36] which is an efficient method for lattices combination [37]. Equal weights are allocated to the systems used in lattices combination.

5. ASR experiments

5.1. Speech corpus

ASR systems are trained and evaluated using Aurora-4 corpus [16]. Aurora-4 is a medium vocabulary task based on the Wall Street Journal (WSJ0) corpus. The multi-condition training set consists of 7137 utterances from 83 speakers. The speech utterances in the multi-condition training set are both clean and noisy. The noisy utterances were created by corrupting clean speech utterances by six different noises (airport, babble, car, restaurant, street, and train) at 10-20 dB signal to noise ratio (SNR). The evaluation set was derived from WSJ0 5K-word closed-vocabulary test set which consists of 330 utterances spoken by 8 speakers. This test set was recorded by a primary Sennheiser microphone and a secondary microphone. 14 test sets were created by corrupting these two sets by the same six noises used in the training set at 5-15 dB SNR. These 14 test sets can be grouped into 4 subsets: clean, noisy, clean with channel distortion, noisy with channel distortion, which will be referred to as A, B, C, and D, respectively. All the data used for the experiments in this paper are sampled at 16 kHz.

5.2. ASR systems

The acoustic models are CNNs consisting of 7 hidden layers, with the first two layers are convolutional layers followed by five fully-connected layers. The convolutional layers used 1-dimension filters, applying convolutions and max-pooling on the frequency axis. Acoustic features for training CNNs are 40-dimensional FBANK features, extracted from 25 ms windows every 10 ms, together with their delta and acceleration coefficients. The features are spliced with 5 frames on each side of the current frame. Utterance-level mean normalization is performed on static features. Multi-condition training data are used. The first convolutional layer has 128 filters of size 8. It is followed by a max-pooling layer with pooling size of 3 and a pooling step of 3, and then, a second convolutional layer. The second convolutional layer has 256 filters of size 4. The output of the second convolution layer is passed to 5 fully-connected hidden layers. Each fully-connected hidden layer has 2048 nodes. The output layer consists of 2298 nodes which are the number of tied triphone HMM states. The state-level alignments for training all the CNN acoustic models are obtained from a speaker adaptive training (SAT) HMM-GMM system, trained on multi-condition training data using MFCCs features. The CNN acoustic models are trained with back-propagation algorithm [38] based on cross-entropy criterion. The task-standard WSJ0 bi-gram language model is used for training and decoding. Experiments are performed using the Kaldi speech recognition toolkit [17].

5.3. Results

WERs of single-stream ASR systems are shown in Tab. 1. Considering the average WERs which are the average of 14 WERs on 14 individual test sets, the system using the $\mathbf{S}_{\mathrm{G_T}}$ weighting provides relative WER reduction of 2.4% and 1.1% compared to the baseline system 1 not using weighting and the baseline system 2 using the EBM spectral masking technique, respectively. When single-stream systems using the $\mathbf{S}_{\mathrm{G_I}}$, $\mathbf{S}_{\mathrm{G_F}}$ and $\mathbf{S}_{\mathrm{G_T}}$ weighting are combined in multi-stream systems, the WERs of the multi-stream systems are reduced (see Tabs. 2 and 3). These WERs are further reduced when the system using the $\mathbf{S}_{\mathrm{O}}$ weighting is included.

Table 1: *WERs of single-stream baseline systems and of systems using ASMs weighting.*

System \ Condition	A	B	C	D	Avg.
Baseline 1 (no weighting)	4.22	7.35	7.94	18.89	12.11
Baseline 2 (using EBM)	3.94	7.32	8.03	18.57	11.95
Weighting using $\mathbf{S}_{\mathrm{G_I}}$	4.04	7.37	7.70	18.88	12.09
Weighting using $\mathbf{S}_{\mathrm{G_F}}$	4.00	7.60	8.28	19.03	12.29
Weighting using $\mathbf{S}_{\mathrm{G_T}}$	3.94	7.22	7.23	18.51	11.82
Weighting using $\mathbf{S}_{\mathrm{O}}$	4.11	7.39	7.98	18.62	12.01

Table 2: *WERs of multi-stream ASR systems using posterior probabilities combination. The weighting techniques used in the combined single-stream systems are separated with commas.*

Combination \ Condition	A	B	C	D	Avg.
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}$	3.87	7.05	7.64	18.18	11.63
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}, \mathbf{S}_{\mathrm{O}}$	3.87	6.99	7.53	18.06	11.55
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}, \mathbf{S}_{\mathrm{O}}$, EBM	3.85	6.98	7.49	17.92	11.48

Table 3: *WERs of multi-stream ASR systems using lattices combination. The weighting techniques used in the combined single-stream systems are separated with commas.*

Combination \ Condition	A	B	C	D	Avg.
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}$	3.89	6.93	7.19	18.16	11.54
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}, \mathbf{S}_{\mathrm{O}}$	3.83	6.87	7.34	18.04	11.47
$\mathbf{S}_{\mathrm{G_I}}, \mathbf{S}_{\mathrm{G_F}}, \mathbf{S}_{\mathrm{G_T}}, \mathbf{S}_{\mathrm{O}}$, EBM	3.81	6.81	7.23	17.91	11.38

Lattices combination seems provide lower WER while having a higher decoding cost than posterior probabilities combination. Relative gains of 5.3% and 4.0% WER are obtained with lattices combination of 4 single-stream systems using three individual ASMs and one overall ASM, compared to the baseline system 1 and the baseline system 2, respectively. This fact suggests that the weighting realized by the individual and overall ASMs are complementary to each other to improve ASR performance. Combining lattices of the systems using the ASMs weighting and that using the EBM technique provides relative gains of 6.0% and 4.8% WER compared to the baseline system 1 and the baseline system 2, respectively.

6. Conclusion

This paper proposed a new method for weighting T-F representation of speech using auditory saliency for noise-robust ASR. The proposed method is inspired from and consistent with the human auditory attention mechanism which weights representation of environment to bias the perception towards salient events [13]. When being applied in multi-stream ASR framework, relative gains of up to 5.3% and 4.0% WER were observed when comparing the multi-stream system using the proposed method with the baseline single-stream system not using weighting and that using the conventional EBM spectral masking technique, respectively. Combining the multi-stream system using the proposed method and the single-stream system using the EBM spectral masking technique reduced further the WER.

7. References

[1] 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, pp. 745–777, April 2014.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[4] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1315–1329, July 2016.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, October 1994.

[6] H. Hermansky and S. Sharma, "Temporal patterns (TRAPs) in ASR of noisy speech," in *Proc. IEEE ICASSP*, Phoenix, Arizona, USA, March 1999, pp. 289–292.

[7] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 3832–3836.

[8] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proc. IEEE ICASSP*, Las Vegas, Nevada, USA, March - April 2008, pp. 4733–4736.

[9] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Proc. INTERSPEECH*, Brisbane, Australia, September 2008, pp. 898–901.

[10] M. R. Schadler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 131, pp. 4134–4151, May 2012.

[11] H. Hermansky, J. R. Cohen, and R. M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, pp. 1968–1985, September 2013.

[12] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Phil. Trans. R. Soc. B*, vol. 372, pp. 1–10, February 2017.

[13] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanism for allocating audtiory attention: An auditory saliency map," *Current Biology*, vol. 15, pp. 1943–1947, November 2005.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, November 1998.

[15] H. Hermansky, "Multistream recognition of speech: dealing with unknown unknows," *Proccedings of the IEEE*, vol. 101, pp. 1076–1088, May 2013.

[16] N. Parihar and J. Picone, *Aurora working group: DSR front end LVCSR evaluation: AU/384/02*. Institute for Signal and Information Processing Technical Report, 2002.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, Hawaii, USA, December 2011.

[18] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. L. Wang, "A direct masking approach to robust ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1993–2005, October 2013.

[19] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1296–1305, August 2014.

[20] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, pp. 2336–2347, April 2009.

[21] A. Narayanan and D. L. Wang, "The role of binary mask patterns in automatic speech recognition in background noise," *J. Acoust. Soc. Am.*, vol. 133, pp. 3083–3093, May 2013.

[22] Z.-Q. Wang and D. L. Wang, "Robust speech recognition from ratio masks," in *Proc. IEEE ICASSP*, Shanghai, China, March 2016, pp. 5720–5724.

[23] J. van Hout and A. Alwan, "A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4105–4108.

[24] J. W. Hall, M. P. Haggard, and M. A. Fernandes, "Detection in noise by spectro-temporal pattern analysis," *J. Acoust. Soc. Am.*, vol. 76, pp. 50–56, July 1984.

[25] J. R. Movellan, *Tutorial on Gabor filters*. http://mplab.ucsd.edu/tutorials/gabor.pdf.

[26] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, pp. 1233–1258, December 1987.

[27] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 1941–1944.

[28] A.-R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 4273–4276.

[29] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[30] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*. Kluwer Academic Publishers, 1994.

[31] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, August 2012.

[32] J. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1741–1752, August 2007.

[33] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP*, Dallas, Texas, USA, March 2010, pp. 4266–4269.

[34] W. Lee, J. Kim, and I. Lane, "Multi-stream combination for LVCSR and keyword search on GPU-accelerated platforms," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 3320–3324.

[35] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proc. IEEE ICASSP*, Hong Kong, April 2003, pp. 741–744.

[36] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, pp. 802–828, October 2011.

[37] F. Xiong, S. Goetze, and B. T. Meyer, "Combination strategy based on relative performance monitoring for multi-stream reverberant speech recognition," in *Proc. IEEE ICASSP*, New Orleans, Louisiana, USA, March 2017, pp. 4870–4874.

[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 9, pp. 533–536, 1986.