INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Full-Sentence Correlation: a Method to Handle Unpredictable Noise for Robust Speech Recognition

Ji Ming, Danny Crookes

Queen's University Belfast, Belfast BT7 1NN, U.K.
j.ming@qub.ac.uk, d.crookes@qub.ac.uk

Abstract

We describe the theory and implementation of full-sentence speech correlation for speech recognition, and demonstrate its superior robustness to unseen/untrained noise. For the Aurora 2 data, trained with only clean speech, the new method performs competitively against the state-of-the-art with multicondition training and adaptation, and achieves the lowest word error rate in very low SNR (-5 dB). Further experiments with highly non-stationary noise (pop song, broadcast news, etc.) show the surprising ability of the new method to handle unpredictable noise. The new method adds several novel developments to our previous research, including the modeling of the speaker characteristics along with other acoustic and semantic features of speech for separating speech from noise, and a novel Viterbi algorithm to implement full-sentence correlation for speech recognition.
Index Terms: speech recognition, noise robustness, full-sentence correlation, unseen/unpredictable noise

1. Introduction

In speech recognition, a major and unsolved challenge is to find a method that can handle unpredictable noise. In deep neural network (DNN) systems, the typical methods used to handle noise include speech enhancement (e.g., [1, 2]), multi-condition training (e.g., [3–6]) aided by data augmentation (e.g., [7–10]), and feature or model adaptation (e.g., [11–15]). These methods share a common characteristic: they require noise data, for noise reduction or feature/model formation. Therefore these methods may be limited for their applicability to handle unpredictable noise, for which data are variable and difficult to predict. Examples include music, human speech, and their combinations, to name a few. These noises usually change fast, so that it may not be easy to predict them with conventional noise estimation (e.g., [16]). Moreover, these noises can vary vastly from time to time and case to case (just think about the sheer variety in music genres), such that it would be difficult to collect enough training data to model all the possible noise conditions. These noises can come from broadcast media (e.g., TV, radio or Internet), and pose a great challenge for accurate home, entertainment or in-car speech recognition.

Like classical speech models based on multicondition training, DNN systems trained for one set of noise may generalize badly to different noise conditions [13, 17] (also see the examples in the experimental section of this paper). Therefore for improving the ability to handle unseen unpredictable noise we need to consider alternative methods. This paper presents such an alternative. We consider direct correlation of very long speech segments, i.e., full speech sentences, for speech recognition. The new method has the potential to significantly improve noise robustness without requiring noise training, and hence could be used to handle untrained or unpredictable noise. We view our method as trying to match directly *wide*-time-range speech, a complement to deep learning. This work extends our previous studies [18] from speech enhancement to speech recognition. The extensions include the refined optimization problem for picking out speech from noise, including the use of speaker characteristics as a constraint, and a novel iterative Viterbi algorithm to implement full-sentence speech correlation for speech recognition.

2. Long segment correlation vs neural networks for noise robustness

The starting point for our new method is the ZNCC (zero-mean normalized correlation coefficient) measure which we use to compare two speech segments to achieve noise-robust matching. For a noisy speech segment and a potential matching clean speech segment their ZNCC is defined as follows:

$$R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) = \frac{\sum_{l=-L}^{L}[x_{t+l} - \mu_{\mathbf{x}}]^{\mathrm{T}}[s_{\tau+l} - \mu_{\mathbf{s}}]}{\sigma_{\mathbf{x}}\sigma_{\mathbf{s}}} \quad (1)$$

where $\mathbf{x}_{t\pm L}$ represents a noisy speech segment centered at frame x_t and consisting of $2L+1$ consecutive frames from x_{t-L} to x_{t+L}; $\mathbf{s}_{\tau\pm L}$ represents a clean speech segment taken from a clean training speech sentence centered at some frame s_τ with $2L+1$ consecutive frames from $s_{\tau-L}$ to $s_{\tau+L}$. In our experiments, we represent each frame by using its power spectral density (PSD) or transform, to be detailed later. In (1), $\mu_{\mathbf{x}}$ stands for the mean frame vector of segment $\mathbf{x}_{t\pm L}$, i.e., $\mu(\mathbf{x}_{t\pm L}) = \sum_{l=-L}^{L} x_{t+l}/(2L+1)$, and $\sigma_{\mathbf{x}}$ stands for the zero-mean Euclidean norm of segment $\mathbf{x}_{t\pm L}$, i.e., $\sigma_{\mathbf{x}}^2 = \sum_{l=-L}^{L}[x_{t+l} - \mu(\mathbf{x}_{t\pm L})]^{\mathrm{T}}[x_{t+l} - \mu(\mathbf{x}_{t\pm L})]$. The same definitions apply to the clean training speech segment $\mathbf{s}_{\tau\pm L}$, with mean frame vector $\mu_{\mathbf{s}}$ and zero-mean Euclidean norm $\sigma_{\mathbf{s}}$.

In [18] we described an oracle experiment, to demonstrate that the ZNCC measure is potentially capable of handling noise without requiring noise training. In the experiment, we included the clean testing speech in the training data, to form the best matching speech segments of arbitrary length up to complete sentences. Given noisy testing speech, we studied under what condition(s) the best matching speech could be found, from all the clean training data, by using ZNCC to compare their segments and to pick out the matching segments based on maximum ZNCC, *without any noise prediction*. Fig. 1 summarizes the experimental results, showing that as the length of the speech segments being compared increases, the probability of finding the best matching speech segments increases rapidly, approaching one, *regardless of the types of noise*. Since this is achieved without noise training, this method, of using ZNCC to compare very long speech segments for speech matching, could be used to handle unpredictable noise. To show that this is a characteristic of the ZNCC measure, we conducted the same oracle experiment by using Gaussian to compare the segments

http://dx.doi.org/10.21437/Interspeech.2019-2127

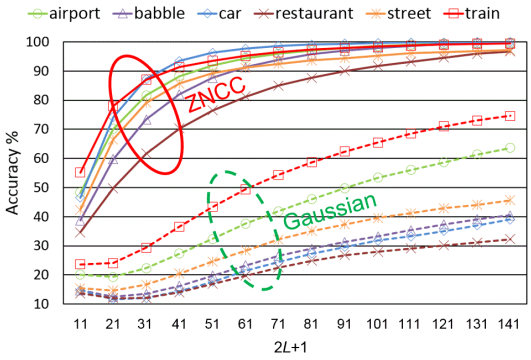

Figure 1: *An oracle experiment showing accuracy of finding best matching speech segments as a function of segment length L (in number of frames) without noise prediction, based on maximum ZNCC (upper group of lines) and maximum Gaussian likelihood (lower group of lines), for six types of noise with SNR = −5 dB, for 57,919 noisy testing speech segments for each type of noise and 1,124,863 clean training speech segments involving 486 speakers based on the TIMIT database.*

and to pick out the matching segments based on maximum likelihood. The results are shown in the lower group of lines in Fig. 1. For the latter, the probability of finding the best matching speech segments does increase with the segment length, however, the increase is unimpressive. In [18], we provided a theory that confirms the generality of the above phenomena.

It is noted that in neural nets similar correlations are performed for the input data (i.e., speech segments). For example, in a typical DNN an input-layer neuron takes a speech segment $\mathbf{x}_{t\pm L}$ and produces a scalar activation output r with the following affine transform before a nonlinear activation function:

$$r(\mathbf{x}_{t\pm L}, \mathbf{w}') = \mathbf{x}_{t\pm L}^{\mathrm{T}}\mathbf{w} + b = \mathbf{x}'^{\mathrm{T}}_{t\pm L}\mathbf{w}' \qquad (2)$$

where $\mathbf{w}'$ is the augmented weight vector (with the bias b). As the input data are usually mean removed and normalized [19], the inner product between the normalized input data and the weights bear a resemblance to the ZNCC calculation. In current DNNs, an input-layer neuron typically takes a *short* speech segment (also called a context window) as input, of a segment length typically below 20 frames. As shown in Fig 1, this just is not long enough to accurately identify the matching speech segments in noise, even with the perfect match included in training, explaining why DNNs require noise training to gain noise robustness. One may wonder if the higher-layer neurons or the neurons in a recurrent DNN perform longer segment matching. The answer is no. Generally, these neurons do not *directly* match longer segments. Instead, they take the lower-layer or earlier-stage short-segment matching activations as input. Since these activations lack noise robustness, the higher-layer or later-stage neurons may just learn inaccurate abstracts.

The aim of this research is to implement the ZNCC to perform direct matching of full speech sentences, i.e., to go as far to the right in Fig. 1 as possible, to maximize the noise robustness assuming no noise prediction. This research is a complement to the deep learning research. By going "wide" (i.e., directly matching very long speech segments), we seek improved noise robustness that is not attainable within the conventional "narrow" (i.e., short speech segments) deep learning. A major obstacle to implementing the sentence-long ZNCC is the lack of enough training data to include the matching estimates of all possible full speech sentences, which are typically hundreds to thousands of frames long. We solve this problem by formulating an optimization problem and solving it by jointly exploiting the acoustic, semantic and speaker-class features of the underlying speech sentences to be matched, detailed below.

3. Sentence ZNCC for speech recognition

Given a noisy speech sentence, without having a matching estimate of the underlying clean speech sentence, we form the estimate by chaining short speech segments chosen from the clean training data. This segment-chain based estimate, while having to achieve maximum ZNCC, must satisfy both acoustic and semantic constraints for being valid speech (or it may just simulate the *noisy* speech instead of the underlying clean speech). Specifically, this estimate must be acoustically sound, with the characteristics of a speaker and regular acoustic patterns of speech; moreover, it must make semantic sense, containing regular subword-word-phrase-sentence patterns of language that separate speech from noise. We encode these requirements into a set of constraining conditions, used to regularize the selection of the short training segments to form the underlying speech sentence estimate that maximizes the ZNCC. We represent a noisy sentence as a sequence of short segments $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$, where $\mathbf{x}_t$ denotes the segment centered at frame x_t consisting of $2L+1$ frames from x_{t-L} to x_{t+L}. For simplicity, we assume a common L being used for all the short segments and so L can be implied in the expression. We may view $\mathbf{x}_t$ as a typical context window modeled by an artificial neuron [see (2)]. It is assumed that there are sufficient training data to catch these short windows' statistics.

Given the noisy speech sentence $\mathbf{X}$, we represent by $\mathbf{S} = (g_1\mathbf{s}_{m_1}, g_2\mathbf{s}_{m_2}, ..., g_T\mathbf{s}_{m_T})$ an estimate of the underlying clean speech sentence based on chaining short training speech segments $\mathbf{s}_{m_t}$, where $\mathbf{s}_{m_t}$ is the training segment as an estimate of the underlying speech segment in noisy segment $\mathbf{x}_t$, with g_t being the gain of the estimate, and m_t being the model vector. In our research, we assume that the training sentences are properly forced aligned and further grouped into speaker classes, each speaker class of sentences sharing similar speaker characteristics. As such, we can group the training segments into state and further into speaker classes within each state. So we define the model parameters that identify each training segment $\mathbf{s}_{m_t}$ by $m_t = (q_t, c_t, n_t; u_t)$, where q_t is the state index of the training segment after its central frame, c_t is the segment's speaker-class index after the training sentence it is from, and n_t is the segment's serial number given its state/speaker group. Additionally, we include the segment's model (word or subword) name for information, indexed by u_t after the segment's central frame. These model parameters will be used to locate a training segment, and to construct the acoustic, semantic and speaker-class constraints for chaining the training segments to form the estimate of the underlying speech sentence. When we find the optimal estimate, we finish speech recognition with the corresponding model name sequence.

3.1. A constrained optimization problem

We seek a matching estimate of the underlying speech sentence by solving a constrained maximization problem:

$$\max_{\mathbf{S}}[\eta(\mathbf{X}, \mathbf{S}) = \log R(\mathbf{X}, \mathbf{S}) + \alpha \log P_{\mathrm{as}}(\mathbf{S}) - \beta E_{\mathrm{sp}}(\mathbf{S})] \quad (3)$$

where $R(\mathbf{X}, \mathbf{S})$ is the full-sentence ZNCC between the noisy $\mathbf{X}$ and a potential estimate $\mathbf{S}$, $P_{\mathrm{as}}(\mathbf{S})$ and $E_{\mathrm{sp}}(\mathbf{S})$ are probabilis-

tic measures of $\mathbf{S}$ being valid speech, and $\eta(\mathbf{X}, \mathbf{S})$ is the objective function to maximize with α and β being the Lagrange constants. By maximizing $\eta(\mathbf{X}, \mathbf{S})$ we jointly maximize the sentence-long ZNCC for finding the matching estimate, and the probability of the estimate to be valid speech. Given limited training data, we assume that this estimate is most likely to be the matching estimate of the underlying speech sentence. We only consider the estimates with $R(\mathbf{X}, \mathbf{S}) > 0$.

The constraint $P_{\text{as}}(\mathbf{S})$ is the probability of a training segment chain $\mathbf{S}$ to be valid speech, which includes both acoustic and semantic conditions modeled as follows:

$$\log P_{\text{as}}(\mathbf{S}) = \frac{1}{T}\Big\{ \log \prod_{t=1}^{T} a[q_{t-1}][q_t] + \sum_{i=1}^{I} \log p_i(d_i) + \sum_{u=1}^{U} \log p_u(d_u) \Big\} \quad (4)$$

The first term in the brackets corresponds mainly to the semantic constraint, with the matrix $a[i][j]$ encoding the probabilities of permissible state transitions within a model and across models (e.g., bigram word language probabilities) of valid speech. In other words, only the chains $\mathbf{S}$ with a state sequence $(q_1, q_2, ..., q_T)$ that fulfill the appropriate phonetic, lexical and language constraints of proper speech can score highly. The second and third terms correspond mainly to the acoustic constrains of proper speech. Assume that the estimate $\mathbf{S}$ has traversed I different states and U different models, with d_i and d_u being the durations spent in each state i and model u, respectively. The two terms calculate the appropriate state and model duration probabilities in a given $\mathbf{S}$, based on the duration probability distributions p_i and p_u learned from clean speech. Only the $\mathbf{S}$ that have an appropriate state sequence $(q_1, q_2, ..., q_T)$ and model sequence $(u_1, u_2, ..., u_T)$ that are similar to what are usually found in clean, valid speech can score the two terms highly. The constraint $E_{\text{sp}}(\mathbf{S})$ in (3) further regularizes the speaker characteristics of the estimate for it to be valid speech. It is the speaker-class entropy of the training segments that form the estimate $\mathbf{S}$. We assume a single, same speaker through out the sentence. Therefore the training segments should come from a single speaker class, or from as few speaker classes as suitable while maximizing the ZNCC. This can be achieved by minimizing the speaker-class entropy as implemented in (3). The entropy is defined as follows:

$$E_{\text{sp}}(\mathbf{S}) = -\sum_{c=1}^{C} h_{\text{sp}}(c) \log h_{\text{sp}}(c) \quad (5)$$

where $h_{\text{sp}}(c)$ is the normalized histogram of the speaker class c in $\mathbf{S}$, based on the associated speaker-class sequence $(c_1, c_2, ..., c_T)$, assuming $1 \leq c_t \leq C$ where C is the number of speaker classes in the training data.

3.2. An iterative Viterbi algorithm

We propose a novel, iterative Viterbi algorithm to solve the constrained maximization problem (3). Unlike conventional Viterbi algorithms which calculate the state path by assuming independent observations in each state, and hence lack noise robustness, we calculate the state path by maximizing, approximately, the constrained, full-sentence ZNCC based $\eta(\mathbf{X}, \mathbf{S})$ in each state, for the noisy sentence $\mathbf{X}$. The new algorithm is outlined below.

We start by obtaining an initial sentence estimate $\hat{\mathbf{S}}^0$ by performing a conventional Viterbi search, by assuming independent noisy segments $\mathbf{x}_t$ in each state and by using the *segmental* ZNCC, defined in (1), as the state-based likelihood for each $\mathbf{x}_t$; in each state, we obtain this likelihood by maximizing over all training segments (n_t) and speaker classes (c_t), with a unit gain g_t for each training segment. Then we update this initial estimate by iteratively performing the novel Viterbi search, by maximizing the full-sentence based $\eta(\mathbf{X}, \mathbf{S})$ as the state-based likelihoods. In the $(k+1)$th iteration ($k \geq 0$), we calculate the likelihood for noisy segment $\mathbf{x}_t$ in state i by using the recursion:

$$\eta(\mathbf{X}, [\hat{\mathbf{S}}_t^{k+1}(i), \hat{g}_{t+1}^k \hat{\mathbf{s}}_{m_{t+1}}^k, ..., \hat{g}_T^k \hat{\mathbf{s}}_{m_T}^k]) = \max_{j, m_t | q_t = i, g_t} \eta(\mathbf{X}, [\hat{\mathbf{S}}_{t-1}^{k+1}(j), g_t \mathbf{s}_{m_t}, \hat{g}_{t+1}^k \hat{\mathbf{s}}_{m_{t+1}}^k, ..., \hat{g}_T^k \hat{\mathbf{s}}_{m_T}^k]) \quad (6)$$

In (6), $\hat{\mathbf{S}}_t^{k+1}(i)$ stands for a partial sentence estimate up to training segment $\hat{\mathbf{s}}_{m_t}$ with state $q_t = i$, in the $(k+1)$th iteration. As shown in (6), to calculate the full-sentence $\eta(\mathbf{X}, \mathbf{S})$ to identify the best $\hat{\mathbf{s}}_{m_t}$ we need the estimates of the remaining training segments $\mathbf{s}_{m_\tau}$ with $\tau = t+1, t+2, ..., T$. But these are not available for the current iteration. Therefore we replace these by their estimates from the previous (the kth) iteration. In (6) we have included the estimation of the gains for the training segments assuming that the optimal gains, that maximizes $\eta(\mathbf{X}, \mathbf{S})$, may vary within a pre-defined range $[g_{\min}, g_{\max}]$. As described in [18], this estimation can be executed efficiently by using a golden section method. In our experiments, we have found that the above iterative Viterbi algorithm converges quickly, typically in three to five iterations from the initial estimates.

4. Experimental studies

As the first stage of evaluation, we use Aurora 2 [20] to test the proposed method. Aurora 2 is designed for speaker-independent recognition of digit sequences in noisy conditions with eight different noise types with variable SNRs from -5 dB to noise free distributed in three test sets: Set A (subway, babble, car, exhibition), Set B (restaurant, street, airport, station), and Set C (subway, street, with a different channel characteristic). Aurora 2 offers two training sets: 1) clean training set containing 8440 sentences from 55 male and 55 female speakers (about 76 sentences per speaker), and 2) multicondition training set containing both clean and noisy sentences for four different noise types that are the same as those in test Set A. In our experiments, we only use the clean training set to build our model without any noise prediction, and we compare our model against the state-of-the-art DNN systems which are all built upon the multicondition training data.

In our proposed method, we represent each frame (25 ms long with a 10 ms period) by using 64-dimensional mel-filterbank power outputs, which are scaled down to the 4th root to suppress their dynamic range. Around each frame, we create an 11 frames long segment (i.e., $2L + 1 = 11$) used for sentence matching. This is similar to the typical context window modeled in artificial neurons for speech processing. Following convention, we model each digit with 16 left-to-right states without state skipping. We group the training sentences into the 110 natural speaker classes for modeling the speaker entropy of a sentence estimate. We further assume that the optimal gain for each training segment varies within the range $[0.25, 4]$. In (3), we assume constant $\alpha = 0.1$ and $\beta = 1$ in all our experiments (we have tested some other values around these and found no significant difference in the results). We use a Gaussian density to model the duration distribution for each state and word.

Table 1: *Aurora 2 word error rate (WER%), averaged over the noise types within each test set, comparing the state-of-the-art DNN methods all trained with multicondition data and/or with adaptation, against the proposed method trained with only clean speech data without any noise prediction (-: Data is not available).*

Method	Set A (SNR (dB))					Set B (SNR (dB))					Set C (SNR (dB))				
	∞	10	5	0	-5	∞	10	5	0	-5	∞	10	5	0	-5
DNN adapt [12]	-	1.1	3.5	13.5	-	-	1.0	4.3	16.8	-	-	1.4	4.2	14.8	-
DRNN [21]	0.9	4.1	10.7	32.9	70.2	0.9	8.6	21.4	51.3	82.9	-	-	-	-	-
MTAE [22]	0.7	2.3	5.0	17.9	53.6	0.7	7.3	20.1	47.9	73.1	0.7	5.4	10.9	36.8	68.3
DDAE [23]	0.6	2.5	5.9	21.0	57.9	0.6	10.7	27.3	58.7	92.0	-	-	-	-	-
CNN [24]	2.4	6.1	15.2	44.9	80.5	2.4	11.0	22.6	52.0	81.3	-	-	-	-	-
PBTRNN [25]	1.4	3.9	9.9	29.9	65.4	-	-	-	-	-	-	-	-	-	-
Tandem [26]	-	2.7	5.4	14.3	-	-	-	-	-	-	-	-	-	-	-
Proposed	0.7	3.6	8.6	15.7	40.2	0.7	3.0	8.5	17.5	41.5	0.6	3.2	8.4	19.8	44.8

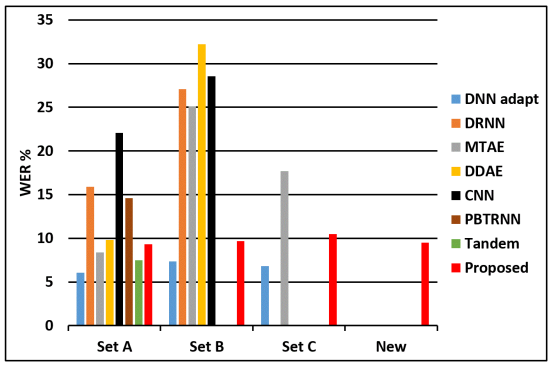

Figure 2: *WER averaged over SNR=10, 5, 0 dB for the Aurora 2 test sets and the new test set containing three new types of noise.*

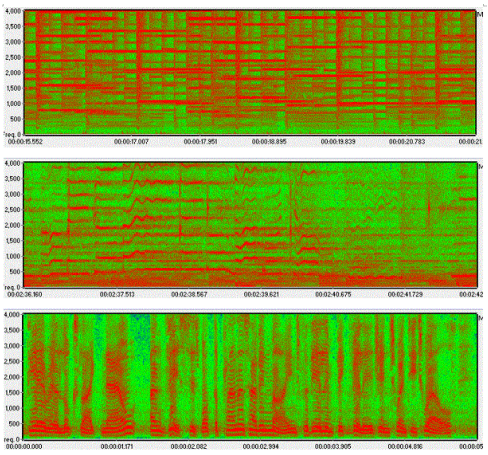

Figure 3: *Three new types of noise (top to bottom): polyphonic ring, pop song, broadcast news.*

Table 2: *WER% for the new noises by the proposed new method.*

Noise	SNR (dB)		
	10	5	0
Polyphonic ring	3.3	5.1	7.0
Pop song	3.2	5.4	12.4
Broadcast news	6.3	13.1	30.0

Table 1 presents the word error rate (WER) of our proposed method, trained with only the clean training set, compared to the DNN methods found in the literature that are all trained with the multicondition data [12, 21–26] and with additional adaptation during testing [12]. For clarity, we present the WER set wise, by averaging the WER across the different noise types within each test set. Fig. 2 provides a more compact view of Table 1, by further averaging the WER in each test set over three SNRs: 10, 5 and 0 dB, for all the methods (subject to the availability of data). As shown in Table 1 and Fig. 2, the proposed method is competitive with the best DNNs, especially in the low SNRs. It achieves the lowest WER in the very low SNR (-5 dB) and has a good generalization ability across the different noise and channel types. It is noted that many of the DNN methods trained on Set A noises generalize poorly on Set B – they have some dramatic WER increase when applied to the mismatched noises.

The purpose of this study is to find a method that can handle unpredictable noise. To deepen this study, we created a new test set by adding three new types of noise to the clean test sentences in Aurora 2. The three new noises are 1) a polyphonic mobile phone ring, 2) a pop song segment with a mixture of background music and the voice of a female singer, and 3) a broadcast news segment from a male speaker. The spectral characteristics of the three new noises are shown in Fig. 3. As can be seen, these types of noise can be highly complex and nonstationary, and can be potentially very different from case to case. Hence it could be difficult to model them with conventional noise estimation, pre-training or adaptation. These can be part of our daily experience and represent some very difficult interferences for speech recognition. Table 2 presents the WER by the new method for the new noises, with the average WER over all the noises and SNR=10, 5, 0 dB shown in Fig. 2 alongside the other test sets. The new method has indeed performed consistently across all these noise types to a competitive recognition accuracy, with only clean speech data training.

5. Conclusions

It is shown that by directly matching full speech sentences, by using ZNCC, it is possible to significantly advance the noise robustness without requiring noise training. The implementation, however, is difficult given limited training data. In this paper we described a method – to formulate the idea as a constrained maximization problem and to solve the problem with an iterative Viterbi algorithm. We have evaluated the new method on Aurora 2 and beyond, and achieved superior performance in terms of robustness to unseen unpredictable noise.

6. References

[1] M. Fujimoto and T. Nakatani, "Feature enhancement based on generative-discriminative hybrid approach with GMMs and DNNs for noise robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5019–5023.

[2] Y.-H. Tu, I. Tashev, S. Zarar, and C.-H. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2531–2535.

[3] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7398–7402.

[4] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 7–19, 2015.

[5] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4375–4379.

[6] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 796–806, 2016.

[7] M. Karafiat, F. Grezl, L. Burget, I. Szoke, and J. Cernocky, "Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2454–2458.

[8] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1469–1477, 2015.

[9] T. Schrank, L. Pfeifenberger, M. Zohrer, J. Stahl, and P. Mowlaee, "Deep beamforming and data augmentation for robust speech recognition: Results of the 4th CHiME challenge," in *Proceedings CHiME*, 2016, pp. 18–20.

[10] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5220–5224.

[11] J. Li, J. T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5537–5541.

[12] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1296–1305, 2014.

[13] V. Mitra, H. Franco, C. Bartels, J. van Hout, M. Graciarena, and D. Vergyri, "Speech recognition in unseen and noisy channel conditions," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5215–5219.

[14] K. H. Lee, W. H. Kang, T. G. Kang, and N. S. Kim, "Integrated DNN-based model adaptation technique for noise-robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5245–5249.

[15] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large scale domain adaptation via teacher-student learning," in *Proceedings Interspeech 2017*, pp. 2386–2390.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, 2001.

[17] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Proceedings ASRU*, 2015.

[18] J. Ming and D. Crookes, "Speech enhancement based on full-sentence correlation and clean speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 531–543, 2017.

[19] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 14–22, 2012.

[20] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings ISCA ITRW ASR*, 2000.

[21] A. L. Maas, Q. V. Le, T. M. ONeil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proceedings Interspeech 2012*, pp. 22–25.

[22] H. Zhang, C. Liu, N. Inoue, and K. Shinoda, "Multi-task autoencoder for noise-robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5599–5603.

[23] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, "Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 350–355.

[24] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings Interspeech 2015*, pp. 11–15.

[25] P. Brakel, D. Stroobandt, and B. Schrauwen, "Bidirectional truncated recurrent neural networks for efficient speech denoising," in *Proceedings Interspeech 2013*, pp. 2973–2977.

[26] X. Xie, R. Su, X. Liu, and L. Wang, "Deep neural network bottleneck features for generalized variable parameter HMMs," in *Proceedings Interspeech 2014*, pp. 2739–2743.