FUSION OF DIVERSE DENOISING SYSTEMS FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Naveen Kumar[1], Maarten Van Segbroeck[1], Kartik Audhkhasi[1], Peter Drotár[2], Shrikanth S. Narayanan[1]

[1]Signal Analysis and Interpretation Analysis Lab (SAIL)
Department of Electrical Engineering,
University of Southern California, Los Angeles, CA 90089
[2]Department of Telecommunications,
Brno University of Technology, Technická 12, Brno
{komathnk,audhkhas}@usc.edu, drotarp@feec.vutbr.cz
{maarten,shri}@sipi.usc.edu

ABSTRACT

We present a framework for combining different denoising frontends for robust speech enhancement for recognition in noisy conditions. This is contrasted against results of optimally fusing diverse parameter settings for a single denoising algorithm. All frontends in the latter case exploit the same denoising algorithm, which combines harmonic decomposition, with noise estimation and spectral subtraction. The set of associated parameters involved in these steps are dependent on the noise conditions. Rather than explicitly tuning them, we suggest a strategy that tries to account for the trade-off between average word error rate and diversity to find an optimal subset of these parameter settings. We present the results on Aurora4 database and also compare against traditional speech enhancement methods e.g. Wiener filtering and spectral subtraction.

Index Terms— Robust Large Vocabulary Speech Recognition, Speech Enhancement, Diversity, ROVER

1. INTRODUCTION

Current state-of-the-art speech recognition systems manage to achieve excellent recognition rates when applied to clean non-distorted speech. However, it is commonly observed that ASR performance degrades rapidly when operating in challenging real-life environments with a decreased signal-to-noise ratio. The presence of noise in the real-world scenarios is almost inevitable and ASR systems have to be able to robustly cope with a wide range of noisy environmental conditions.

A common approach to deal with the robustness issue in ASR, is to add a speech enhancement unit in the front-end to improve the spectral quality of recorded signal. Most popular methods of speech enhancement at the signal level include Spectral Subtraction [1], [2], Minimum Statistics [3] or Wiener Filtering based methods [4]. In this paper, we select a noise compensation method based on harmonic decomposition and spectral subtraction [5]. The denoising algorithm first estimates the noise from the residual part of the input signal obtained after removing the periodicity caused by voiced speech. The enhanced signal is then obtained by subtracting spectrum of the noise estimate from the input speech. The advantage of this denoising method is that it neither requires a speech activity detector nor assumes that important prior knowledge of noise is available, e.g. the stationarity of the noise over a relatively large time window. Moreover, it is able to reduce unwanted speech degradations by the limited leakage of voicing energy in the spectral subbands of the noise prior to subtraction.

The disadvantage of such a denoising algorithm, and one can easily generalize this over similar noise reduction methods, is that it can be very sensitive to parameter settings, thus requiring extensive parameter tuning on a development set. Moreover, it is often difficult to obtain a single parameter setting that works for all noise types and conditions. Hence, there have been attempts to combine results from multiple speech enhancement systems. For example, in [6] authors try an ensemble learning approach for fusion of time-frequency masks at the signal level. In this paper, we propose fusion of multiple diverse denoising algorithms at the hypothesis level for robust speech enhancement in ASRs. We show that each denoising algorithm generates diverse hypotheses which can be combined together using ROVER [7] to obtain an improvement in Word Error Rate (WER). In addition, we show that a diverse combination of different parameter settings for a single denoising algorithm can also yield an improvement in WER.

Results obtained on denoising experiments conducted on the Aurora4 dataset establish the algorithm's robustness to different noise types. In addition, we show that a further improvement in recognition accuracy can be obtained by combining the outputs from different systems using ROVER [7], which exploits the diversity between them.

The paper is organized as follows: Section 2 explains the basic concept of the speech enhancement technique and the denoising parameters involved. In Section 3 the dataset and setup of the KALDI speech recognition toolkit are discussed. Section 4 outlines the inter and intra-system fusion approaches. We discuss the problem for finding the optimally diverse subset of parameter settings for the intra-system fusion case. Experiments and results are presented in Section 5. Concluding remarks are given in Section 7.

2. SPEECH DENOISING ALGORITHM

In this work, we consider a bandlimited, noisy speech signal $x(t)$ which is assumed to be the sum of clean speech signal $s(t)$ and additive noise $n(t)$. Both the noise and speech are assumed to be uncorrelated non-stationary signals. In order to obtain denoised signal $y(t)$ we implement the speech denoising algorithm introduced in [5]. The denoising involves three main processing steps: voicing

removal, noise estimation and spectral subtraction.

The first step is to remove periodicity from the noisy speech signal. To obtain an unvoiced signal, the noisy speech is decomposed into its harmonically related components. Here, an initial pitch estimate is computed by the subharmonic summation method [8]. A pitch synchronous framing is then subsequently applied to the signal to obtain overlapping segments with the length of two pitch periods and a single period of frame shift. The unvoiced noisy signal is computed as a per frame subtraction of the input speech signal and harmonic component of the input signal, estimated as a sum of sine and cosine terms of harmonics of the instantaneous pitch. Multiplication by a linear modulation factor further takes into account the change in amplitude over the duration of the samples. Detailed description of estimation of modulation parameter and pitch frequency can be found in [5] or [9]

The short-time sub-band energy of the noise can be estimated from the minimum statistics of the short-term Fourier spectrum of $x(t)$. The minimum statistics noise power spectral density estimation approach is based on tracking minima of a short term power spectral density estimate in frequency sub-bands [10]. This approach prevents the subtracting of high energy unvoiced speech regions present in $x(t)$ and eliminate the requirement of a voice activity detector.

In order to obtain a denoised version of the noisy signal, the subtraction rule proposed in [10] is adopted. The spectral magnitudes of the noise estimate are subtracted from the spectrum of the noisy signal, taking into account an over-subtraction factor that is computed as a function of the signal-to-noise ratio per frequency sub band. For ASR purposes, features can directly be extracted from the denoised spectrum. Our implementation also foresees a reconstruction of the denoised speech signal in the time-domain after applying an inverse Fourier Transform on each frame taking into account the (unaltered) phase of the noisy signal and a division by the values of the Hamming window.

2.1. Denoising parameters

The result of the denoising procedure is influenced by several parameters that can be modified in order to adapt to observed noise conditions. Therefore, in order to build a noise robust denoising module, it is preferable to find a fixed set of parameters that provide useful speech enhancement under different noise scenarios. In other words, instead of tuning parameters for a specific noise type, we would like to select parameters that provide good overall performance. Another alternative approach that we later adopt in this paper is to combine hypotheses from diverse parameter settings to improve the noise robustness of our system. We discuss a method to select such a subset of parameters in Section 4.

Some important parameters in noise tracking include: size of running window for minimum statistics noise estimation(*vsm*), noise floor level of aperiodic signal (*nfl*),the noise reduction factor during speech frames (*rfs*) and the noise factor to compensate for introduced non-linear effect (*nsa*), optimized using a grid search of possible values. Specifically, we searched over the grid (*vsm, nfl, rfs, nsa*) defined by the product of the sets *vsm*= $[10, 20, 40]$, *nfl*= $[0.05, 0.1, 0.2]$, *rfs*= $[0.25, 0.5, 1.0]$ and *nsa*= $[0.05, 0.1, 0.125]$. The rest of denoising parameters such as the noise reduction factor during noise frames, upper and lower noise level of voicing were kept constant during denoising procedure. We choose 23 different parameters settings $p_1, \ldots, p_{23}$ from this grid for tuning and diverse parameter selection experiments.

3. METHODOLOGY

3.1. ASR Setup using KALDI

For our ASR experiments we use the Kaldi [11] open-source Speech Recognition Toolkit. Kaldi was built on top of OpenFst [12] libraries, with the aim to be flexible, easy to understand, and to provide extensive Weighted Finite State Transducer (WFST) and math support.

While training an acoustic model, Kaldi makes use of an update based strategy using previous alignments generated by a simpler model for training a more complicated model. For example, a monophone training model is first trained using a flat alignment (i.e. uniform priors). Once trained, the alignments generated by this model are used to train a triphone model. The process is repeated for several training steps to obtain a better acoustic model using standard MFCC $+\Delta + \Delta\Delta$ features. A brief description of the different models trained for our experiments is given in Table 1 below.

System	Context	Transformation	Alignment
mono0a	Monophone	CMN	flat
tri1	Triphone	CMN	mono0a
tri2a	Triphone	CMN	tri1
tri2b	Triphone	CMN+LDA+ MLLT	tri2a
tri3b	Triphone	CMN+LDA+ MLLT+SAT	tri2b

Table 1. Different systems trained using Kaldi. The final results presented in Section 5 are obtained using system *tri3b*. *(CMN- Cepstral Mean Normalization, LDA- Linear Discriminant Analysis, MLLT- Maximum Log-Likelihood Transformation, SAT- Speaker Adaptive Training)*

3.2. Aurora4 Dataset

Our recognition experiments are presented on the Aurora4 database containing utterances derived from the WSJ0 Wall Street Journal 5k-word dictation task [13] under SNR levels ranging from 5db to 15db. The test data set comprises 7 sets of read speech from 8 different speakers, artificially corrupted by additive noise after applying a P.341 filtering characteristic [14] to clean data samples at 16 kHz. We conduct our experiments only on the close talking microphone signal giving us a total of 330 test samples for each of the noise types : no noise (set 01), car (set 02), babble (set 03), restaurant (set 04), street (set 05), airport (set 06) and train (set 07). The clean training set contains 7138 utterances from 83 speakers, also otherwise known as the *SI_84* subset of the WSJ0 dataset. We additionally use a development set containing 1206 utterances from 10 speakers to search for an optimally diverse subset of parameter settings for the HD denoising algorithm. Details of this intra-system fusion strategy are presented in the next section.

4. FUSION OF DIVERSE SYSTEMS

The fusion of multiple denoising algorithms is done at the level of ASR hypotheses. We use the Recognition Output Voting Error Reduction (ROVER) scheme [7] to perform a weighted combination of the one-best hypothesis from each system. The weight parameter for ROVER is tuned on the dev set. For our experiments we use $\alpha = 0.7$. This shall be henceforth referred to as inter-system fusion since it involves fusion of hypotheses from ASRs with different denoising frontends viz. Harmonic Decomposition (HD), Wiener

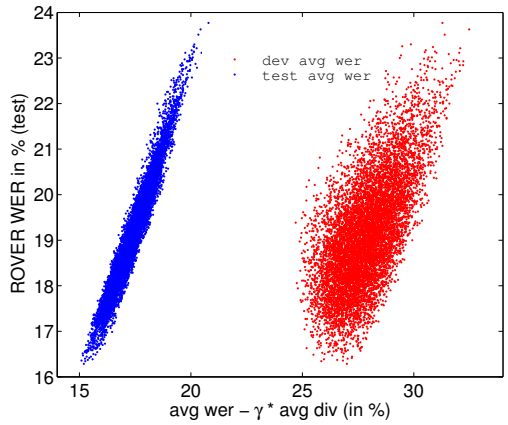

Fig. 1. Scatter plot showing the relation between ROVER WER on the test set, average WER and average diversity across datasets. γ is estimated on the dev set.

Filtering (WF), Cepstral Mean Normalization (CMN) and Spectral Subtraction (SS) [15].

We also perform intra-system fusion experiments in addition to the above inter system fusion. We combine hypotheses generated from different parameter settings of the HD algorithm. Selecting the best subset of parameter settings for this purpose is an NP-hard problem and requires a large amount of computation. We thus exploit the ROVER WER-diversity decomposition approach, presented by Audhkhasi et. al in [16]. They study the relation between diversity of ASR systems being fused and the ROVER WER. Let $E(r,h)$ be the WER or Levenshtein distance between the reference transcription r and the hypotheses h. Let h^* be the ROVER fusion of K hypotheses $\{h_1,\ldots,h_K\}$. Then [17] shows that the ROVER WER approximately decomposes as

$$\underbrace{E(r,h^*)}_{\text{ROVER WER}} \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K} E(r,h_k)}_{\text{Average WER}} - \gamma \underbrace{\frac{1}{K}\sum_{k=1}^{K} E(h^*,h_k)}_{\text{Diversity}} . \quad (1)$$

The parameter γ is tuned on the dev set. We use the WER computed on the development set as a proxy for the average system WER shown in Eqn.(1). However, the diversity term for each fusion subset can be computed on the test set directly since it does not require reference transcriptions.

Fig.1 shows the relation between ROVER WER, average WER and average diversity for each subset of parameters being fused. Each point in the plot corresponds to ROVER fusion of a subset to 4 parameters settings of the HD algorithm. As expected, the ROVER WER has a stronger correlation in the oracle case when the average WER is directly computed on the test set (blue scatter plot). The red scatter plot shows the case when the average WER on the dev set in instead used as a proxy. Fig. 1 thus suggests that the average WER and diversity can be used to roughly estimate the fusion WER and hence choose the best subsystem of parameters.

5. EXPERIMENTAL RESULTS

We first present some initial results in Table 2 using clean train and the test sets mentioned earlier. The sequential training procedure described in Section 3 is applied and for each test set we perform a decoding using each of the systems in Table 1. The *tri3b* system, which trains triphone models and performs LDA and MLLT transforms of the features, along with Speaker Adaptive Training (SAT), achieves the best WER consistently. Henceforth, all presented results are trained using this best system.

The ASR results corresponding to each denoising algorithm are presented in Table 3. The extracted features for all denoised audio streams, are first mean normalized (CMN). Hence, the baseline numbers are obtained using normalization of the features only and are denoted by CMN. The best parameter setting for the HD denoising algorithm shown in the last row is tuned on a small subset of the test set comprising first 10 utterances by number from each of the 8 speakers (80 utterances in all). It is important to mention here that this is done only to adapt the HD denoising algorithm to the acoustic conditions present in the test set, as the denoising algorithm doesn't really make use of any speaker/ sentence information. We select the best parameter setting among $\{p_1,\ldots,p_{23}\}$ based on the average WER across all noise types. In addition, since the HD algorithm is known to produce non-linear artifacts, the clean condition training data was also preprocessed by the implemented method to match test conditions.

Results indicate that the implemented denoising algorithm (HD) is comparable in performance to the baseline algorithms on most of the noise types. The HD denoising algorithm in fact provides an improvement in performance on the non-stationary noise types like "babble" (03) and "train" (07). The degradation in performance on the "clean" (01) and "car" (02) test sets, can be attributed to the inherent distortion caused by the denoising algorithms and the fact that the parameters for our method were tuned to give a low overall WER. This leads to the important observation that with the choice of a single optimal parameter setting the implemented system is unable to adapt to all noise types at once. This motivates our next approach for fusion of diverse denoising front-ends.

5.1. Fusion of diverse denoising front-ends

The systems HD$^{\delta_k}$ in Table 4 denote different intra-system ROVER fusions for diverse parameter settings of the HD algorithm. HD$^{\delta_0}$ corresponds to the first oracle system with the best ROVER WER $E(r,h^*)$ for 4 parameter settings on the test set. The WER used for each system is the average for different noise types.

For the HD$^{\delta_1}$ system, we use average of the WERs due to each individual system, computed directly on the test set. The average WER is then used to estimate the ROVER WER according to Eqn.(1). The system HD$^{\delta_2}$ uses average WER computed on the dev set instead, for selecting the best 4 parameter subsystem. To put things in perspective, we also present the performance of the HD algorithm in the first row, with the parameter settings tuned on the

system \ test	01	02	03	04	05	06	07	Avg.
mono0a	13.8	33.5	49.5	57.6	54.9	47.7	55.3	44.6
tri1	6.4	17.6	31.9	40.5	39.6	30.8	40.3	29.6
tri2a	5.7	18.1	32.6	40.7	39.0	30.2	39.5	29.4
tri2b	5.5	16.0	37.0	46.1	42.1	35.5	41.8	32.0
tri3b	4.2	9.4	24.1	31.2	25.9	21.4	28.1	20.6

Table 2. Word Error Rate (WER) obtained using clean training noising and only Cepstral Mean Normalization of the features.

Aurora4, 16kHz, clean condition training.								
	Close Talk							
Test	01	02	03	04	05	06	07	Avg.
CMN	**4.2**	**9.4**	24.1	31.2	25.9	**21.4**	28.1	20.6
WF	4.4	9.7	20.9	26.6	23.0	22.4	23.6	18.7
SS	4.5	10.8	22.9	24.9	22.9	23.0	24.0	19.0
HD*	6.6	12.6	**17.0**	**24.1**	**22.7**	22.7	**19.2**	**17.8**

Table 3. Word error rate (in %) on the Aurora4 test sets using the proposed and baseline denoising algorithms. HD* corresponds to the HD denoising algorithm with parameters tuned on a small subset of the test set.

dev set instead.

The last two rows in Table 4 show the WER obtained by word frequency based ROVER (wfR) and confidence weighted ROVER (cfR) of all denoising algorithms (CMN, WF, SS, HD). For the HD algorithm we use the same parameter settings as the HD* system (Table 3 Row 4). Here, the noisy transcriptions corresponding to the four denoising methods are combined to achieve a further reduction in WER with respect to the use of a single ASR system. This probably hints at the complementary nature of each system allowing ROVER to combine their strengths. The best average WER of 14.2% on Aurora4 test set 01-07, is comparable to the performance obtained by state-of-the-art (and more complex) noise compensation methods.

6. DISCUSSION

The intra-system ROVER fusion results in Table 4 show that there is merit to combining hypotheses from ASRs driven by different parameter settings of a single denoising algorithm. The systems HD$^{\delta_k}$ are able to improve the performance of the HD algorithm on the clean (01) and car (02) test sets, by fusing results from other parameter settings that perform better on these noise types. In addition, the performance of the HD$^{\delta_1}$ system is comparable to the best intra-system ROVER result (HD$^{\delta_0}$). This corroborates the diversity-decomposition based strategy proposed for selecting the optimal subset of parameter settings. The results of HD$^{\delta_2}$ exhibit a similar trend, in spite of a mismatch between the dev set and the test set average WERs. This mismatch is owing to the dev set vocabulary containing out-of-vocabulary verbalized punctuations such as the one shown in the example below.

id: (050o0209)

```
REF: finally ,COMMA japanese business is moving
to meet the demand .PERIOD
HYP: finally COMMON japanese business is moving
to meet the demand PERIOD
```

Clearly, the proposed denoising system based on ROVER fusion of diverse parameter settings is more versatile compared to the HD denoising system based on a single extensively tuned parameter setting (Table 4 Row 1). The disadvantage for the latter system results from its tendency to over-adapt to the noise conditions in the dev set. Hence, even the system HD$^{\delta_2}$ based on an automatically found subset of parameter settings improves the performance over this optimally tuned system based on a single parameter setting.

Although, diverse combination of different parameter settings helps, we observe that the inter-system ROVER fusion systems outperform the intra-system ROVER fusion systems. This can be attributed to be the inherent lack of diversity among chosen candidate parameter settings. A choice of parameter settings keeping their diverse combination in mind might be a better choice and is an open problem.

Aurora4, 16kHz, clean condition training.								
	Close Talk							
Fusion	01	02	03	04	05	06	07	Avg.
HD	7.3	18.9	24.5	29.6	28.6	28.5	25.1	23.2
HD$^{\delta_2}$	5.9	13.5	19.1	25.9	24.2	26.0	19.9	19.2
HD$^{\delta_1}$	5.7	10.7	16.0	22.8	20.3	21.5	17.5	16.4
HD$^{\delta_0}$	5.5	10.9	16.3	22.5	19.5	21.7	17.6	16.3
wfR	4.1	7.9	16.9	22.1	19.0	18.0	18.8	15.3
cfR	3.6	6.9	16.1	21.2	18.2	16.4	18.6	14.2

Table 4. WER reported on the Aurora4 test set by combining different denoising algorithms using ROVER. Combining outputs from each system using ROVER consistently improves the WER in all test sets compared to the best performing individual system.

7. CONCLUSIONS

We present a framework for combining hypotheses from different denoising algorithms in order to achieve robust ASR performance in noisy conditions. We show that even in a single system case, ROVER fusion of speech enhancement with diverse parameter settings can be used to create diversity and obtain robustness to different noise types. We propose a strategy to find an optimally diverse subset of such parameter settings.

Results on the Aurora4 database indicate that the proposed method for combining different denoising algorithms is robust across different noise types and conditions. The obtained WER is comparable to baseline systems and fusion of different parameter settings is useful when dealing with a dataset of unknown or mixed noise type.

In the future, we would like to substitute the diversity computation with an approximate metric thereby eliminating the need for computing the ROVER WER for all parameter subsets on the dev and test sets.

8. ACKNOWLEDGMENTS

This research was supported by the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF). Peter Drotár was supported by the project CZ.1.07/2.3.00/30.0005 of Brno University of Technology.

9. REFERENCES

[1] Yang Lu and Philipos C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, June 2008.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Joon-Hyuk Chang, "Noisy speech enhancement based on improved minimum statistics incorporating acoustic

environment-awareness," *Digital Signal Processing*, vol. 23, no. 4, pp. 1233 – 1238, 2013.

[4] S.V. Vaseghi and B.P. Milner, "Noise compensation methods for hidden markov model speech recognition in adverse environments," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 11–21, 1997.

[5] Maarten Van Segbroeck and Shrikanth Narayanan, "A robust frontend for ASR: Combining denoising, noise masking and feature normalization," in *Proceedings of the IEEE*. IEEE, 2013, pp. 803–806.

[6] Jonathan Le Roux, Shinji Watanabe, and John R Hershey, "Ensemble learning for speech enhancement," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[7] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[8] Dik J Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, pp. 257, 1988.

[9] Hugo Van Hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–213.

[10] R. Martin, "Noise power spectral density estimation based on omptimal smoothing and minimum statistics," *Speech and Audio processing, IEEE Transactions on*, vol. 9, no. 4, pp. 504 –512, July 2001.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[12] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Implementation and Application of Automata*, pp. 11–23. Springer, 2007.

[13] N Parihar and J Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, 2002.

[14] Simão Ferraz De Campos Neto, "The ITU-T software tool library," *International journal of speech technology*, vol. 2, no. 4, pp. 259–272, 1999.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, 1979, vol. 4, pp. 208–211.

[16] Kartik Audhkhasi, A Zavou, P Georgiou, and S Narayanan, "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, March 2014.

[17] Kartik Audhkhasi, Andreas Zavou, Panoyiotis Georgiou, and Shrikanth Narayanan, "Empirical link between hypothesis diversity and fusion performance in an ensemble of automatic speech recognition systems," in *Proc. Interspeech*, 2013.