Improving the Noise Robustness of Prominence Detection for Children's Oral Reading Assessment

Kamini Sabu, Kanhaiya Kumar, Preeti Rao
Dept. of Electrical Engineering
IIT Bombay, India
kaminisabu@iitb.ac.in, kanhaiyanathani8@gmail.com, prao@ee.iitb.ac.in

***Abstract*—Reading skill is a critical component of basic literacy. We aim to develop an automated system to assess oral reading skills of primary school children (learning English as a second language) that could eventually be valuable in the scenario of teacher shortage typical of rural areas in the country. This work focuses on the rating of prosody, an important aspect of fluency in speech delivery. In particular, a system for the detection of word prominence based on prosodic features is presented and tested on real-world data marked by background noise typical of the school setting. To counteract the observed drop in prominence classification accuracy, two distinct approaches to noisy speech enhancement are evaluated for various types of background noise. A recently proposed Generative Adversarial Network(GAN) based method is found to be effective in achieving noise suppression with low levels of speech distortion that minimally impact prosodic feature extraction. The implementation and training of the GAN system is discussed and insights are provided on its performance with reference to that of classical spectral subtraction based enhancement.**

***Index Terms*—Prosody, Prominence, Noisy Speech Enhancement, Generative Adversarial Network(GAN)**

I. Introduction

Reading is a major aspect of literacy skills. Proficient reading skills help in a student's long-term education and further build confidence. As per reading research till date, fluent reading is considered to comprise of word decoding accuracy and prosody. Prosody is the supra-segmental aspect of speech which also has linguistic functions. Speakers tend to focus the listener's attention on the most important parts of the message through use of prosody [1] and is often essential in order to resolve possible ambiguities in the meaning of some utterances. Phrasing is mainly indicative of the ability of a speaker to divide a sentence into meaningful chunks, while prominence indicates the important words (usually words containing new information) by giving special emphasis on these words. In students' reading evaluation task, proper use of prominence acts as a cue to the student's interpretation and understanding of the text. This can thus help in the automatic assessment of student's comprehension [2].

Prominence is acoustically indicated through changes in pitch, intensity and duration of the associated word, the contribution of each of the attributes being dependent on the language. Prominent words are perceived louder and higher-pitched with syllable nucleus lengthening. Besides, [3] shows that there is some spectral tilt observed in words perceived as prominent. There is a lot of research work on detecting prominent words for automatic prominence annotation of transcribed speech databases. Pitch and intensity contours change in unexpected ways during the occurrence of prosodic event and this unpredictability is used by [4] for prominent word detection. RMS energy, RFC(rise-fall-connection) model of f0 contour, syllable nucleus duration and spectral emphasis in 0.5-2k band are considered promising features for prominence of syllables [3]. Considering the fact that prominence also depends on lexical content like type of word, lexical features can accompany acoustic features [5]. Statistical model or classifier approaches with supervised or unsupervised training are commonly used. Typical unsupervised approach is used with duration, intensity, spectral intensity and pitch patterns as features for prominent word detection in Switchboard telephonic corpus [6].

All the research work related to prominence detection till date is focused on adult speech. Our work deals with students' read speech. The task becomes difficult since the target class is L2 language reading by young learners. We found in our database that many students tend to give prominence to function words too. This scenario reduces the importance of lexical features. We therefore restrict ourselves to word-level prosody features to estimate prominence.

For our students' reading evaluation task, students of age group 10-14 years are recorded as they read short English stories. Their native language is a dialect of Marathi. We can clearly observe the effect of native accent on their English speaking styles. This forms a major challenge in the automatic reading evaluation. Further, the scenario being the school environment, around 80% of the data is found to be noisy with major proportion being babble, rain, wind and childPlaying. Babble and childPlaying are known to be the most challenging noises comprising various sounds of impulsive and non-stationary characteristics. The presence of noise is expected to degrade the performance of the assessment system. This problem underlines the need for good speech enhancement system which will act as a preprocessing step for the assessment task. The speech enhancement should ideally obtain noise suppression with minimal speech distortion.

There are many speech enhancement techniques which may or may not require prior noise estimates. Conventional methods rely on spectral subtraction, which uses local information

Authors acknowledge support by MeitY, Govt. of India through Visvesvaraya PhD scheme.

978-1-5386-1224-8/18/$31.00 ©2018 IEEE

to find the noise spectral estimate. The system performance highly depends on way the noise has been estimated. Over estimation distorts the actual speech and under estimation retains noise. Other classical techniques are Wiener filtering, subspace based algorithms and statistical model-based methods. Neural network based enhancement methods, especially the recurrent neural network (RNN or LSTM) based denoising architecture, are being adopted in recent days [7]. In [8], the noise estimates are also concatenated with the input features of deep neural network to improve the performance further. Current speech enhancement techniques operate on the spectral domain and/or exploit some higher-level feature [9]. Most of the methods work on the short-time Fourier spectrum [10], where only spectral magnitude is modified as it is often claimed that short-time phase is not important for speech enhancement. However recent studies [11] have shown that there is significant improvement in speech quality when clean phase spectrum is available.

There has been a large increase in interest towards generative models recently. As the name suggests, these models tend to create the data that is much similar to the data that we provide to them. Generative Adversarial Network(GAN) is considered as a breakthrough in the deep learning field. It has done a wonderful job in the field of computer vision by generating the most realistic images. GAN has been used for speech enhancement, directly on the raw audio waveform [9]. The use of GAN-based enhancement has not been previously investigated in the context of noisy speech recognition or prosody detection however.

Studies related to prominence detection have typically been restricted to clean speech datasets. Some noisy datasets have been used with attention to developing noise-robust features for prominence [6]. Some similar studies in emotion recognition domain exploit multi-condition training [12] to obtain better matching of input speech to the trained acoustic models. However use of enhancement as pre-processing to prominence detection has not been considered. In this paper, we perform prominence classification on children's read speech. We check which features are used by the students in order to give emphasis, if given at all. Noisy conditions are obtained by artificially adding real noise of various types to clean data. Finally, we check the effect of noise as well as that of speech enhancement on prominence detection results.

Next section discusses the prominence detection system used by us. Section 3 describes classical and GAN based speech enhancement methods. Next, datasets and experiments are presented followed by observations and discussion.

II. Prominence Classifier

The overall system discussed is as shown in the Fig. 1. The audio signal is passed through ASR to get word-level alignments. The alignments are used to segment utterance into words. Prosodic features are calculated for every word using pitch, intensity and spectral balance contours. These prosodic features are input to the prominence classifier which gives binary output as to whether the word is prominent or non-prominent. Further, in order to obtain possible improvements, we try speech enhancement of the audio using one of the enhancement techniques (GAN based method and classical spectral subtraction based method).

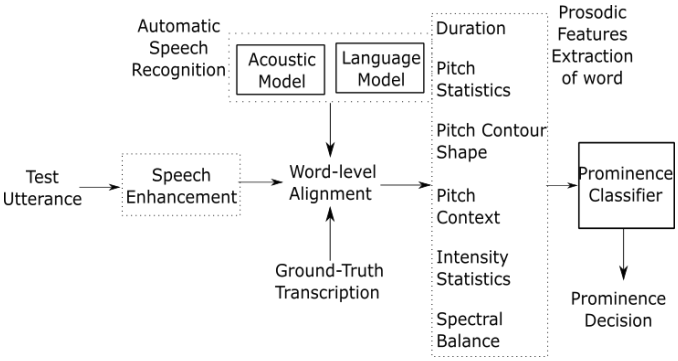

Fig. 1. System Block Diagram

Based on the previous work on prominence detection by [6], [3], [5], [21], [22] we compute following features at word-level:

- Average syllable duration, latency, subsequent pause duration, flags indicating occurrence of pause (silence greater than 150ms) before and after the word
- Spectral balance in sonorant band 0.3-2.3k, Spectral energy in bands - 0-0.5K, 0.5-1K, 1-2K, 2-4K
- Short time energy statistics - Mean, min, max, span, standard deviation, median, RMS Energy on energy contour evaluated at 10ms intervals. All are normalized by the total energy of the utterance energy. Peak intensity is also calculated in dB and normalized by subtracting the mean intensity of utterance.
- Statistical features of pitch - Mean, min, max, span, standard deviation, median. All are z-score normalized at the utterance level. Pitch is computed using Praat toolkit at 10ms hop duration.
- Correlation of pitch contour with ideal rise, fall, peak and valley contours and with Gaussian contour of same length and five different variances - 0.2,0.5,1.0,2.0,5.0
- Correlation of pitch contour in +/-1 word context with ideal rise, fall, peak and valley contours of the same length. Average pitch difference between the adjacent words is also calculated.

Required word-level alignments are obtained on clean audios through forced alignment with ground truth transcription using ASR decoder [23], [24]. The same alignments are used for extracting prosodic features of noisy and enhanced data too. This will help us to study the performance of prominence classification separately from the effect of noisy speech alignment.

The Random Forest tree is found to give the best performance for prominence classification in earlier works [5], [22]. Therefore, random forest tree from WEKA toolkit is used as a prominence detection classifier. Classification training

is supervised with two class labels, prominent and non-prominent.

III. Speech Enhancement

A. Classical methods for speech enhancement

There are many speech enhancement methods which assume that the prior estimate of noise is available, for instance, it is crucial for Wiener filtering [13], for estimating noise covariance matrix in the subspace algorithm [14]. The noise estimates are generally calculated (or updated) in the silence portion which further requires a voice-activity detection (VAD) algorithm. These methods work well when noise is stationary, but fail in more realistic environments (non-stationary noises). Several noise-estimation algorithms have been proposed which update the noise estimates continuously over time [15], [16], [17], [18]. These algorithms propose a method for estimating the noise spectrum based on tracking the minimum of the noisy speech over a finite window. However, these methods are sensitive to outliers and also the noise updates are dependent on the length of the minimum-search window. For our analysis, we have used [19] as the representative of classical methods, whose algorithmic flow is:

- The smoothed power spectrum of noisy speech is computed using the first-order recursive equation.
- Finding the local minima for each bin of power spectrum and updating it by continuously averaging past spectral values.
- Computing the element wise ratio of power spectrum and local minima estimate to find speech presence probability. A varying threshold has been used to classify each of the frequency bins into speech present/absent.
- Further smoothening of speech presence probability has been done across time and frequency.
- The noise spectrum estimate is calculated using this speech presence probability. And then enhanced signal is obtained by subtracting the noise spectrum estimate from original spectrum.

We have reproduced the results claimed by them using an available implementation and the same have been used for the comparison of results in our investigation. Next section describes a more recent speech enhancement technique based on Generative Adversarial Network (GAN).

B. Generative Adversarial network (GAN)

Deep learning generative models are considered most powerful methods for machine learning. Generative Adversarial Network is the extended version of it. It combines two neural networks together and makes them compete against each other so that they can train and improve themselves. It can also be said that both neural networks complement each other by sharing their knowledge to improve themselves. This gives rise to a feedback loop of continuous improvements without human intervention.

1) GAN training process: It can be thought of as a conventional neural network training where G (the generative model) tries to minimize the loss function emulated by D (the discriminator model). Instead of relying on the conventional loss functions whose minimization generally gives blurry results [20], the loss function automatically gets learn satisfying a high-level goal. The process of this adversarial training is also called as a minimax game between G and D. Here, the generative model (G) is trying to learn a mapping from a prior distribution $p_z(z)$ (generally taken to be a random distribution) to the data distribution $p_{data}(x)$, $G : z \rightarrow x$. But, in the enhancement task we need a mapping from noisy signal $\overline{x}$ to the clean signal x. To do so, conditional GAN [20] is used where it learns the mapping from x_c (it is a compressed vector of noisy signal $\overline{x}$) and random noise vector z, to x, $G : \{x_c, z\} \rightarrow x$. GAN's overall objective function is generally defined by,

$$\min_G \max_D V(D,G) = \mathbb{E}_{x,x_c \sim p_{data}(x,x_c)}[\log D(x,x_c)] + \mathbb{E}_{z \sim p_z(z), x_c \sim p_{data}(x_c)}[\log(1 - D(G(z,x_c),x_c))] \quad (1)$$

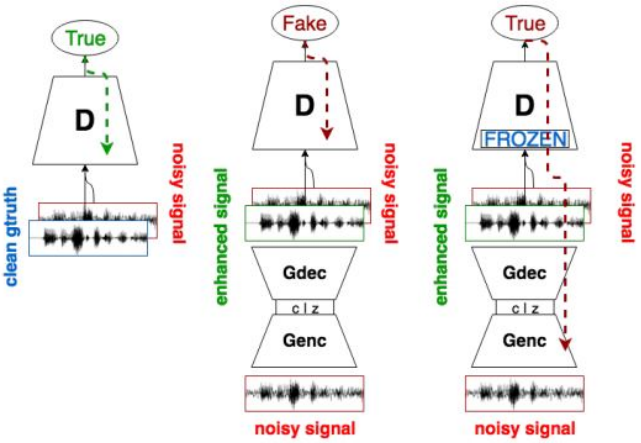

Fig. 2. GAN training process [9]

Fig. 2 summarizes the entire training process. First, the D network is trained via back-propagation using a batch of real pair, composed of a noisy signal ($\overline{x}$) and a clean signal (x). Then, D back-props a batch of fake pair, composed of a noisy signal ($\overline{x}$) and an enhanced signal ($\hat{x}$) that come from G, and classifies them as fake. Finally, D's parameters are frozen and G back-props to make D mis-classify. Dashed lines represent gradient back-propagation.

2) Network Architecture: The G network (in Fig. 2) is similar to the encoder-decoder network. At the bridge point the random noise vector z is concatenated. The encoder is made fully convolutional so that there is no dense layer at all. This enforces the network to focus on temporally-close correlations in the input signal and throughout the whole layering process [9]. There are additional connections between the corresponding encoding and decoding layers similar to U-Nets. Parametric ReLUs have been used as the activation

function after each layer. The D network is similar to encoder network of G, it is also fully convolutional.

IV. Dataset

A. Training Dataset

For the task, we have audio recordings of fluent students as they read short stories in a relatively quiet room using voice recording software on Android tablet at 16kHz sampling rate. From these, 50 perceptually clean audio recordings with over 30dB estimated SNR were selected. All audio lengths vary from 30 to 60 seconds. Some small noisy regions, if any, were removed.

An expert rater (graduate in Linguistics) was asked to listen to the audio recordings and label each of the words in the recording as prominent or non-prominent. The word was to be rated as prominent if rater feels that the student wants to highlight the word as important. Final count of words in the dataset is 5250 of which 1031 are prominent and 4219 are non-prominent. Expert ratings show that function words (especially pronouns and some prepositions and conjunctions) too are given prominence by students in many cases. This may be due to the fact that the students are new to English language and are not able to make distinction between important content words and regular words during reading. This data is used for training the prominence detection classifier. We found that the rater was able to rate prominence with $\approx 95\%$ consistency in artificial noisy situations obtained by adding realistic noise to a 10-story sample of these stories. This means human rater is comfortable with the data distortion even at 10dB SNR and so should be our automated evaluation system.

B. Testing Dataset

We have recordings of students reading short stories in karaoke form using LETS App [25] on Android tablet. These recordings are in school environment and contain background noise. A set of 40 relatively clean audio recordings has been selected for testing. Some small noisy regions, if any, are replaced with silent regions in the same recording. These are divided into two sets of 20 stories, one set is used for training enhancement network, while other is used as testing data.

Testing data is segmented into utterances. Each utterance is labeled by rater for prominence at 3-point scale (no prominence at all, some words prominent, but not as expected, all and only the expected words are made prominent). The expected prominent words are obtained from the narrator audio of corresponding story on the LETS App. Only those utterances which have been rated as 3 by the rater are selected for prominence testing since we know exactly which word was made prominent by the student. There are 50 utterances comprising total 465 words (328 non-prominent and 137 prominent words) in the test data for the prominence classifier.

C. Noise Audios

Different instances of four noise types (rain, wind, babble, childPlaying), comprising half hour each, were collected from the internet [26], [27], [28], [29], [30] and our dataset. Each noise type has its characteristic properties. Rain is broadband with spectral characteristics similar to white noise. Babble refer to the speech-like sounds from many distinct speakers in the background. Wind noise is observed to have no harmonicity, but is energetic. School noise is highly dynamic with highly varying spectral characteristics. Babble and rain are observed to be present over the complete recording. Wind and childPlaying, on the other hand, are intermittent.

V. Experiments

There are 12 different instances of each noise type of which one is chosen randomly and a randomly selected segment from that instance is added to the clean recordings of testing dataset in sectionIV-B at 10dB and 20dB SNR. Noisy versions were created for all the 20 clean recordings used for testing prominence classifier. The desired 465 words are then extracted and used for prominence testing. Therefore, we have 465 words in total for every noisetype-SNR pair in test data. Further, enhancement is also applied on these recordings and features are extracted for the desired words.

For training speech enhancement system, noisy versions are synthesized corresponding to the other set of testing dataset in sectionIV-B of 20 relatively clean audio recordings by adding noise instances, as discussed earlier, accounting for around 10 hours of data. The enhancement task is acting as a preprocessing step for prominence detection task, therefore performance of enhancement task is analyzed through results of prominence detection.

The GAN network implementation is taken from SEGAN [9] and has been trained using the noisy data and their corresponding clean versions. It has been trained for around 50 epochs with RMSprop and learning rate of 0.0002, using batch size of 400. A chunk of waveforms with a sliding window of length one second (16384 samples) with 50% overlapping is used while training. But during testing, the overlap is taken to be zero. For both training and testing, a pre-emphasis filter of coefficient 0.95 is applied to all input samples (during testing, output is correspondingly deemphasized) [9].

We tested it for a noisy audio, and found that it completely removes the noise in the silence portion, but in the speech portion it is slightly distorting the true speech also. We observed that the training data has lot of silence portion. Because of this, the Discriminator network might be distinguishing the clean and noisy utterances based only on the silence content. It may not be able to focus on the noise content in the speech portion. So, we reduced the silence portion in the training data using the available silence locations for the audios. After training GAN with the modified training data, we can observe significant improvement in the speech quality via noise reduction in the speech part in addition to the silence part.

For classical method, we have used the same parameter settings as in [19]. We had also generated some of the results claimed by them on our dataset using the same parameter settings.

The prominence detection classifier is trained on clean data. Testing is performed in four scenarios – clean data, noisy data with four noise types, noisy data enhanced by GAN method and by classical method.

VI. Results and Discussion

Fig. 3 shows the audio waveform, spectrogram and pitch contour of clean, noisy and both the enhanced versions of a typical audio. We can see that GAN performs really well as compared to the classical method. The pitch is nicely recovered which is crucial for the prosody assessment task. Prominence detection classifier performance is evaluated in

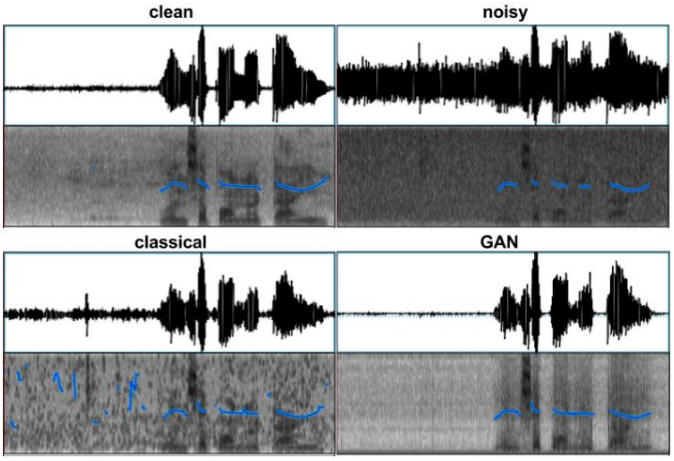

Fig. 3. Waveform, spectrogram and pitch (blue contours) of clean, noisy and enhanced versions of an audio

terms of equal error rate (EER). Table I shows EER values (in %) in different test conditions. Training and testing on clean gives area under the RoC curve 0.794 at 65% precision and 70% recall at EER (equal error rate). When the noisy data is

TABLE I
Equal Error Rate (EER) (in %) for prominence detection classifier in different test conditions (EER=28.3% for clean test data) when word-boundaries from forced alignment on clean data are used

	Noisy		GAN		Classical	
	10dB	20dB	10dB	20dB	10dB	20dB
rain	30.2	29	28.6	29	31	29.3
wind	29.6	29.2	29.2	28.6	30	29
babble	31.1	28.9	30.7	30.2	28	28.9
childPlaying	31.7	30.5	29.2	29	29.9	31.1
white	33.2	32	32.6	31.1	29.3	32.6

fed to the prominence classifier, the performance degrades as can be seen from the table I. Irrespective of the noise type, the degradation is more in 10dB than 20dB. Degradation observed is more on babble and childPlaying noise types as is evident from the dynamic nature of these noise types. Further babble 20dB is less affected than 10dB. However, childPlaying is always affected. GAN has performed well on childPlaying noise at 10dB too.

We can see that enhancement using GAN improves the performance significantly (1%) for 20dB SNR conditions with slight improvement in 10dB SNR. Classical method performed badly compared to GAN based method on all the noises especially 20dB. This might be the case since classical method subtracts local minima, thus overestimating noise in high SNR cases. Performance has significantly improved for babble and white noise in 10dB case, where noise estimate gets evaluated properly. The degradation in performance is especially because pitch contour does not get evaluated properly for audios enhanced using classical method, as evident from Fig. 3. GAN, on the other hand, tends to maintain the pitch contour intact.

TABLE II
Equal Error Rate (EER) (in %) for prominence detection classifier in different test conditions (EER=28.3% for clean test data) when the word-boundaries are obtained from forced alignment on the test data

	Noisy		GAN		Classical	
	10dB	20dB	10dB	20dB	10dB	20dB
rain	31.7	30.5	26.5	28.3	29.9	28.7
wind	32.6	29.6	28.3	28.9	28.7	26.5
babble	28.9	28.6	29.3	26.8	31.4	29.7
childPlaying	33.2	28.3	28.9	28.3	29.8	28.7
white	29	29	29.6	29.6	31.1	28

Table II shows the EER values when word-level alignments used for obtaining the prosodic features are obtained by force aligning the test audio itself instead of using the alignments from clean audio. This is more general case since we won't have clean versions available for actual noisy data. We see that classical method performs better compared to noisy version in this case, but GAN based method outperforms. The results are found to be even better than when alignments from clean version of the data were used. This may be because enhancement introduces the wrong pitch values at the word-silence junctures. These are getting avoided through new word alignment and hence the word-level prosodic features are improving. However, in case of babble, performance degrades after enhancement. This is unusual and needs further investigation. Similar trend has been observed in case of white noise, which may be because pitch is less affected by white noise.

The overall results show that GAN based enhancement is suitable for prominence detection on noisy data. We further want to improve the prosodic features in order to get better prominence detection rate. Even in GAN based method, the performance does not reach performance of clean test data. Some cases of failure involve errors in pitch tracking especially at the juncture of word and silence. Since pitch is very important feature in prominence detection, pitch tracking algorithm should be robust to noise and enhancement. Training on the enhanced data instead of clean data needs to be

References

[1] M. Beckman and J. Venditti, "Tagging prosody and discourse structure in elicited spontaneous speech", in Proc. of Science and Technology Agency Priority Program Symposium on Spontaneous Speech, 2000

[2] P. Schwanenflugel et al., “Becoming a Fluent Reader: Reading Skill and Prosodic Features in the Oral Reading of Young Readers”, in Journal of Educational Psychology, 96(1), 2004, 119–129

[3] F. Tamburini, “Prosodic prominence detection in speech”, in Proc. of International Symposium on Signal Processing and its Applications, 2013

[4] S. Kakouros and O. Rasanen, “Automatic detection of sentence prominence in speech using predictability of word-level acoustic features”, in Proc. of INTERSPEECH, 2015

[5] T. Mishra, V. Sridhar and A. Conkie, “Word prominence detection using robust yet simple prosodic features”, in Proc. of INTERSPEECH, 2012

[6] D. Wang and S. Narayanan, “An acoustic measure for word prominence in spontaneous speech”, in IEEE Transactions on Audio, Speech, and Language Processing, 15(2), 2007, pp. 690–701

[7] 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 Proc. of INTERSPEECH, 2012, pp. 2225

[8] A. Kumar and D. Florencio, “Speech enhancement in multiple-noise conditions using deep neural networks, in Proc. of INTERSPEECH, 2016, pp. 37383742

[9] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network”, arXiv:1703.09452v3, 2017

[10] P.C. Loizou, “Speech Enhancement: Theory and Practice”, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[11] K. Paliwal, K. Wojcicki, and B. Shannon, “The importance of phase in speech enhancement, Speech Communication, vol. 53, no. 4, 2011, pp. 465 494

[12] F. Eyben, F. weninger, B. Schuller, “Affect recognition in real-life acoustic conditions - a new perspective on feature selection”, in Proc. of INTERSPEECH, 2013

[13] J.S. Lim and A.V. Oppenheim, 1978. “All-pole modeling of degraded speech”. IEEE Trans. Acoust. Speech Signal Process. ASSP 26 (3), 197210

[14] Y. Ephraim, H.L. Van Trees, 1993. “A signal subspace approach for speech enhancement”, Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Process. II, 355358

[15] D. Malah, R. Cox, A. Accardi, 1999. “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary environments”, Proc. IEEE Internat. on Conf. Acoust. Speech Signal Process., 789792.

[16] R. Martin, 2001. “Noise power spectral density estimation based on optimal smoothing and minimum statistics”, IEEE Trans. Speech Audio Process. 9 (5), 504512

[17] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging”, in IEEE Trans. Speech Audio Process. 11 (5), 2003, 466475

[18] G. Doblinger, “Computationally efficient speech enhancement by spectral minima tracking in subbands”, in Proc. Eurospeech 2, 1995, 15131516

[19] S. Rangachari, P.C. Loizou, “A noise-estimation algorithm for highly non-stationary environments”, Speech Communication, 48, 2006, 220231

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks, ArXiv: 1611.07004, 2016

[21] A. Rosenberg, E. Cooper, R. Levitan and J. Hirschberg, “Cross-language prominence detection”, in Proc. of Speech Prosody, 2012

[22] G. Christodoulides and M. Avanzi, “An evaluation of machine learning methods for prominence detection in French”, in Proc. of INTERSPEECH, 2014

[23] P. Swarup, “Acoustic Model Training and Adaptation for Children’s Read Speech Recognition”, MTP report, Dept. of Electrical Engineering, IIT Bombay, 2017

[24] H. Tulsiani, “Language Modeling & Miscue Detection for Children’s Read Speech Assessment”, MTP report, Dept. of Electrical Engineering, IIT Bombay, 2017

[25] P. Rao, P. Swarup, A. Pasad, H. Tulsiani, G. Das, “Automatic assessment of reading with speech recognition technology”, in Proc. of International Conf. on Computers in Education, 2016

[26] “Youtube”, https://www.youtube.com, youTube:NoiseAudios

[27] “Nature Sounds for Me”, https://naturesoundsfor.me/, nature Sounds Database

[28] “RWTH Aachen University”, https://www.iks.rwth-aachen.de/en/research/tools-downlaods/wind-noise-database/, wind Noise Database

[29] “University of Toronto”, https://tspace.library.utoronto.ca/handle/1807/66306, The Natural Sound Library, 2014

[30] D. Sahi, “Essential Indian Sound Effects”, Parvati Pictures Pvt Ltd, 1999