Investigation of Neural Network Approaches for Unified Spectral and Prosodic Feature Enhancement

Wei-Cheng Lin*, Yu Tsao*, Fei Chen✦ and Hsin-Min Wang†
*Research Center for Information Technology Innovation, Academic Sinica, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw
✦Department of Electrical and Electronic Engineering, Southern University of Science and Technology, China
E-mail: fchen@sustc.edu.cn
†Institute of Information Science, Academic Sinica, Taiwan
E-mail: whm@iis.sinica.edu.tw

***Abstract*—Most speech enhancement (SE) systems focus on the spectral feature or raw-waveform enhancement. However, many speech-related applications rely on other features rather than the spectral features, such as the intensity and fundamental frequency (f0). Therefore, a unified feature enhancement for different types of features is worth investigating. In this work, we train our neural network (NN)-based SE system in a manner that simultaneously minimizes the spectral loss and preserves the correctness of the intensity and f0 contours extracted from the enhanced speech. The idea is to introduce an NN-based feature extractor to the SE framework that imitates the feature extraction of Praat. Then, we can train the SE system by minimizing the combined loss of the spectral feature, intensity, and f0. We investigate three bidirectional long short-term memory (BLSTM)-based unified feature enhancement systems: *fixed-concat*, *joint-concat*, and *multi-task*. The results of the experiments on the Taiwan Mandarin hearing in a noise test dataset (TMHINT) demonstrate that all three systems show improved intensity and f0 extraction accuracy without sacrificing the perceptual evaluation of the speech quality and short-time objective intelligibility scores compared with the baseline SE system. Further analysis of the experimental results shows that the improvement mostly comes from better f0 contours under difficult conditions such as low signal-to-noise ratio and nonstationary noises. Our work demonstrates the advantage of the unified feature enhancement and provides new insights for SE.**

I. INTRODUCTION

Various speech-related applications are currently flourishing. Ensuring the performance of these applications in a variety of noise environments has become an urgent need. As a pre-processing step in the speech input, speech enhancement (SE) techniques have been widely applied in speech-related applications such as the noise-robust automatic speech recognition (ASR) [1, 2, 3, 4], assistive listening [5, 6, 7, 8], speech coding [9, 10], and speaker verification [11, 12] systems. Recently, neural network (NN) models have been introduced as a fundamental model for the SE task [13, 14]. Owing to their good nonlinear mapping capability, NN-based SE methods have achieved outstanding performance. Numerous works apply different NN models to SE, such as the deep fully connected NN [15], deep denoising autoencoder (DDAE) [16], recurrent NN [17, 18], long short-term memory (LSTM) [17, 18], and convolutional NN [19, 20]. For most NN-based SE works [16, 17, 19], the speech waveform is first converted to a spectral-feature sequence, and a noisy-to-clean mapping process is carried out in the spectral domain. Accordingly, the main objective of these SE tasks is to minimize the distance between the enhanced and clean spectral features.

The success of an SE system depends on whether it can provide benefit to noise-robust speech-related applications. However, most speech applications rely not only on the spectral features but also on other features such as the prosodic features. For a voice conversion (VC) system, in addition to the spectral features (e.g., mel-cepstral coefficients), fundamental frequency (f0) and aperiodicity are important features that need to be considered [21, 22]. *Tranter et al.* [23] reported that in a prototypical speaker-diarization system, accurate voice activation detection (VAD) and gender recognition can yield high diarization performance, and they rely on the intensity and f0 features [24]. In a speech emotion-recognition task, prosodic features such as energy, f0, and duration usually play an important role in the recognition accuracy [25].

In addition, many studies have confirmed the effectiveness of incorporating prosodic features to provide complementary information for enhancing the main task of interest. In the study by *Lin et al.* [26], a hierarchical prosodic model was constructed to facilitate accurate tone information to improve the Mandarin ASR. *Ghannay et al.* [27] combined different word embeddings with prosodic features for ASR error prediction. For pathological voice detection (e.g., Parkinson), prosodic features are also considerably beneficial in the final diagnosis accuracy [28]. The aforementioned works have confirmed the importance of prosodic features and indicated that when the SE module is used as a front-end processor for speech-related applications (in addition to the pursuit of precise spectral mapping), preserving the correctness of the prosodic features should be considered.

According to our literature survey, no prior works have designed an SE system that aims to simultaneously enhance the unified acoustic features (e.g., spectral and prosodic features). In [29], a multi-objective learning approach for SE was proposed to obtain more accurate estimations of the log-power spectral features with the help of a secondary task,

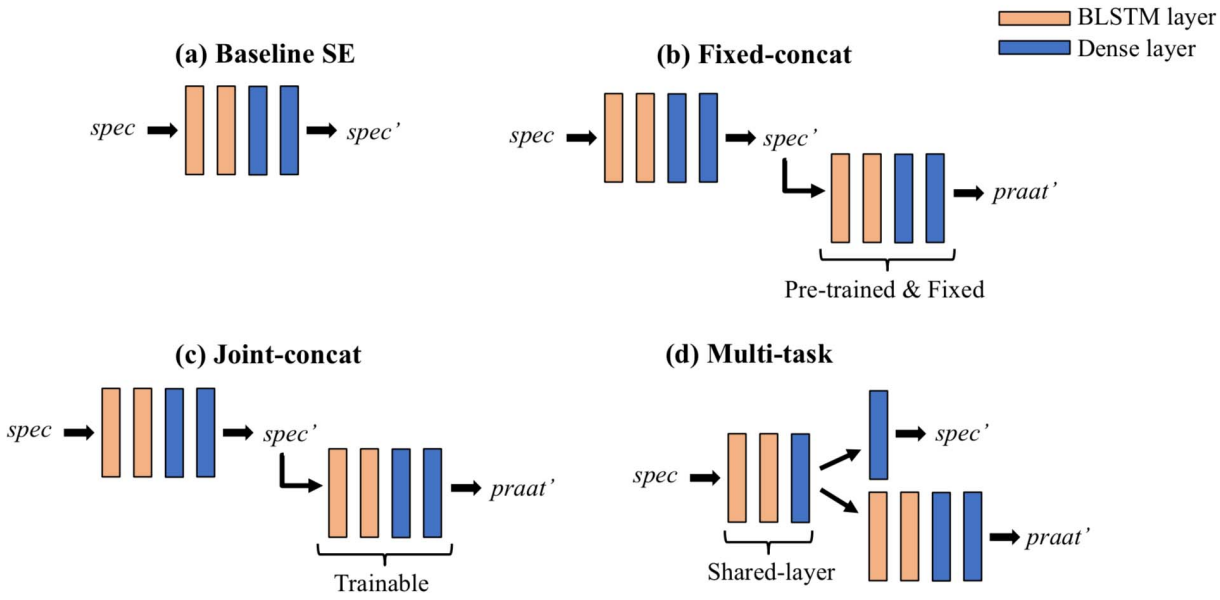

Fig. 1: Overview of different learning approaches.

which considered other spectral features or categorical information. That work could be considered as the closest to our current study. However, the motivations and methods are different. In the current work, we study three SE systems (i.e., *fixed-concat*, *joint-concat*, and *multi-task*) that generate not only enhanced spectral features but also more accurate prosodic features. To evaluate the proposed systems, in addition to the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), which are two standardized evaluation metrics for SE performance, we also measure the correctness of the intensity and f0 contours in terms of the Spearman correlation coefficient. Our experimental results confirm that all three SE methods improve the intensity and f0 extraction accuracy without sacrificing the PESQ and STOI scores compared with the baseline SE system. We note that in contrast to the works that extracted the prosodic features from noisy speech signals [30, 31], the main objective of the unified feature SE methods is to generate speech waveforms with enhanced spectral and prosodic properties. We believe that the enhanced speech waveforms can provide higher speech intelligibility to individuals with normal or impaired hearing [32], especially for tonal languages.

The remainder of this paper is organized as follows. Section II introduces the database used in this study. Section III presents the research methodology. Section IV details the experimental results and analysis. Section V concludes the paper with discussions and future work.

II. Database and Features

A. Database

The speech corpus used in this study consisted of 2,560 Mandarin utterances recorded by eight native speakers (four males and four females). The script for recording was based on the Taiwan Mandarin hearing in a noise test (TMHINT) [33]. A total of 320 unique sentences were provided in which each sentence contained 10 Chinese characters. Each speaker recorded all 320 sentences. The length of each utterance was approximately from 3 to 4 seconds. All utterances were recorded in a quiet environment at a 16-kHz sampling rate.

To prepare the training set, we selected the first 200 utterances of six speakers (three males and three females) as the clean training data. Then, we randomly sampled 100 utterances from the 200 utterances by each speaker and artificially mixed the sampled utterances with 40 noise types from [34] at a signal-to-noise ratio (SNR) that ranged from −10 to 20 dB. This process resulted in $100 * 6 * 40$ noisy training utterances with their corresponding clean references.

In the testing set, all conditions were different from the training set, including different scripts, speakers, and noise types. We selected the last 120 utterances of the remaining two speakers (one male and one female) and mixed the selected utterances with two stationary (i.e., engine and car idle) and two nonstationary (i.e., street and babble) noises at five SNR levels (i.e., −10, −5, 0, 5, and 10 dB). These noise types were not presented in the training set. Overall, the testing set consisted of $120 * 2 * 4 * 5$ noisy utterances.

B. Feature Extraction

The starting and ending silent portions of each clean utterance (and the corresponding noisy ones) were discarded using a pitch-based VAD. For the spectral feature extraction (FE), we conducted a short-time Fourier transform (STFT) of 512 sample points (i.e., 32-ms frame size) for every 256 sample points (i.e., 16-ms frame shift), which resulted in a sequence of frame-based spectral features of 257 dimensions for each utterance. The prosodic features, including the intensity (dB) and f0 (Hz), were extracted using the Praat toolbox [35] using

TABLE II
PROSODIC FEATURE ENHANCEMENT RESULTS OF THE BASELINE AND MULTI-TASK SE APPROACHES UNDER DIFFERENT GENDER, NOISE TYPE, AND SNR CONDITIONS (ALL P-VALUES ARE LOCATED BETWEEN 0.01 AND 0.05)

	Intensity						**f0**					
	Gender		*Noise type*		*SNR*		*Gender*		*Noise type*		*SNR*	
Approach	Female	Male	NonStat.	Stat.	Low	High	Female	Male	NonStat.	Stat.	Low	High
Noisy	0.765	0.783	0.757	0.791	0.525	0.941	0.335	0.411	0.371	0.376	0.109	0.534
BaselineSE	0.805	0.867	0.815	0.857	0.722	0.912	0.633	0.511	0.556	0.588	0.417	0.675
Multi-task	0.817	0.874	0.824	0.866	0.738	0.917	**0.675**	0.518	**0.584**	0.609	**0.457**	0.690

TABLE I
TESTING RESULTS OF DIFFERENT SE APPROACHES (ALL P-VALUES ARE LOCATED BETWEEN 0.01 AND 0.05)

	SE		**Intensity**	**f0**
Approach	*PESQ*	*STOI*	*ρ*	*ρ*
Noisy	1.532	0.692	0.774	0.374
BaselineSE	1.978	0.737	0.836	0.572
Fixed-concat	1.966	0.730	0.833	**0.586**
Joint-concat	1.953	0.733	0.841	**0.591**
Multi-task	1.976	0.739	0.845	**0.597**

the same frame rate and frame size as the spectral feature. Note that because we first adopted VAD and then used a smooth f0 contour in our work, the f0 contour was continuous (e.g., Fig. 2(d)). Some research works apply such continuous f0 processing, such as in multimodal emotion recognition [36]. As a result, after FE, each frame was represented as a 259-dimensional vector (257 spectral (denoted as *spec* hereafter) and 2 prosodic (denoted as *praat* hereafter because they were extracted using Praat) dimensions).

III. METHODOLOGY

Because our goal was to compare the different learning approaches for a unified feature enhancement, we fixed the base NN architecture used in all methods. The base NN model was composed of two fully connected bidirectional LSTM (BLSTM) layers (i.e., the number of nodes was the same as the dimension of the input features), followed by two dense layers (the first with 300 nodes and the second with the same number of nodes as the target feature dimensions), using the Leaky ReLU activation function. All models were trained using the RMSprop optimizer to minimize the custom loss between the output and target features, which is defined as

$$Loss = a * MSE(spec) + b * MAE(praat) \quad (1)$$

where a and b are weighting parameters, *spec* and *praat* represent the spectral and prosodic features, respectively, and *MSE* and *MAE* denote the mean square and mean absolute errors, respectively. These structures and parameters are empirically determined.

A. Learning Approaches

In this work, we compared three unified SE methods with the normal SE method. The overview of these methods is shown in Fig. 1. We describe the details in this section.

1) *Baseline SE*: This is a normal SE model trained to output the enhanced *spec′* feature from the input noisy *spec* feature. Because it does not use the *praat* feature information, the loss function is equivalent to that in Equation (1) with [a, b] set to [1, 0].
2) *Fixed-concat*: Fig. 1(b) shows that this method directly concatenates a pre-trained FE model after the SE model to feed back the losses of the predicted intensity and f0 contours to train the SE model. The pre-trained FE model is fixed during the SE model training. In this work, we trained the FE model using all 1,200 clean speech utterances from the training set. The loss in the FE model training was the *MSE* of the predicted prosodic features based on the input clean *spec* feature with respect to the corresponding clean *praat* targets extracted by the Praat toolbox from the clean waveform. We could consider the FE model as an alternative Praat toolbox that operated on the spectral feature rather than the raw waveform. Therefore, it could be concatenated after the SE model to generate the *praat* feature loss information to train the SE model. To achieve this goal, the performance of the FE model must be sufficiently high. The evaluation of the FE model is discussed in Section IV-A.
3) *Joint-concat*: Figs. 1(b) and 1(c) show that the only difference between the *joint-concat* and *fixed-concat* methods is that the SE and FE models are jointly trained in *joint-concat*, whereas the pre-trained FE model is fixed during the SE model training in *fixed-concat*. Hence, we can expect that the FE model will better fit with the SE model in *joint-concat*. The FE model in the *joint-concat* system can be regarded as a prosodic feature extractor that is trained to generate prosodic features from the enhanced spectral features.
4) *Multi-task*: One common learning approach to utilize diverse information is the multi-task learning. The architecture of the *multi-task* system in this study is shown in Fig. 1(d). The main concept is to build a common feature representation (a shared layer) for the SE and FE tasks. Thus, effective regularization is induced, and more accurate outputs are generated for both tasks.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Settings

We report the results of the testing set mentioned in Section II-A. As presented earlier, all the SE systems used the same NN model structure (i.e., 2-BLSTM and 2-Dense). Therefore, we could fairly compare their SE performance. We use the

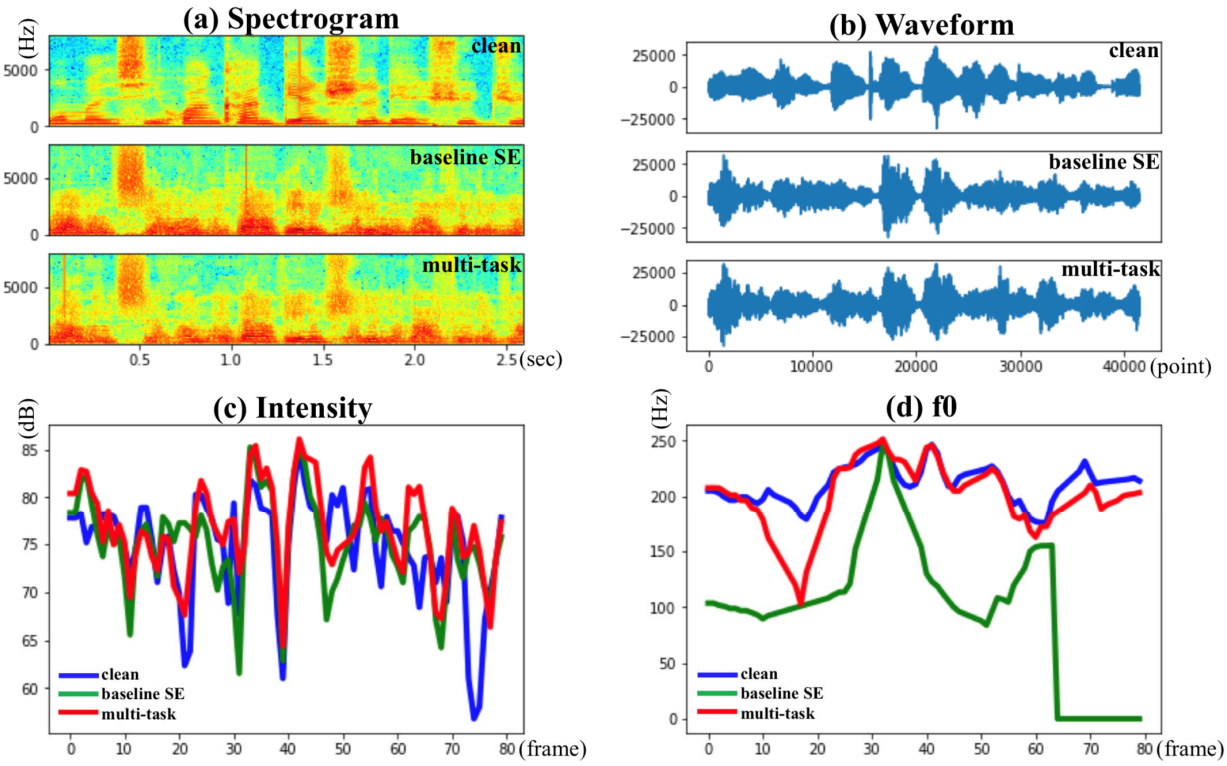

Fig. 2: Example of different enhanced features using the baseline and multi-task SE approaches, including the spectrogram, waveform, and intensity and f0 contours extracted by the Praat toolbox from the enhanced speech. The results of the clean speech are also listed for comparison.

inverse STFT to convert the enhanced spectral output of each SE model into waveform and evaluate the SE performance in terms of PESQ [37] and STOI [38]. To evaluate the output intensity and f0 features, we compared the feature contours extracted by Praat from the enhanced waveform and the corresponding reference clean waveform in terms of the Spearman correlation coefficient (ρ). The weighting parameters [a, b] in Equation (1) were set to [10, 0.1] in all methods (i.e., *fixed-concat*, *joint-concat*, and *multi-task*). By testing the 240 clean utterances from our testing set, the pre-trained FE model used in *fixed-concat* achieved $\rho = 0.935$ and $\rho = 0.820$ for the intensity and f0 contour extractions, respectively, compared with the contours extracted by Praat. The performance was considered sufficient for use in the *fixed-concat* SE system.

B. Results and Analysis

All p-values (two-sided for a hypothesis test whose null hypothesis expresses that the two sets of data are uncorrelated) of the Spearman correlation coefficient (ρ) in Tables I and II were located between 0.01 and 0.05, which were considered as statistically significant. Table I lists the summary of the test results of the different SE approaches. From Table I, we first note that all the evaluation metrics (PESQ, STOI, and Spearman correlation coefficient ρ for intensity and f0) were rather poor under a noisy condition. In particular, the accuracy of prosodic FE was seriously deteriorated by the noise. Next, we note that the baseline SE system could already improve the PESQ and STOI scores with notable margins. Meanwhile, the prosodic features extracted from the enhanced waveform by the baseline SE system also yielded higher Spearman correlation coefficients than those extracted from the noisy waveform.

When investigating the three unified feature enhancement systems, we first note that all systems could notably enhance the accuracy of the f0 estimation (*joint-concat* and *multi-task* could also produce enhanced intensity estimation) while maintaining the PESQ and STOI scores compared with the baseline SE system. Among the three systems, the *multi-task* system achieved the best performance in all the evaluation metrics. In the next process, we used the *multi-task* system as a representative to illustrate the advantage of considering the unified features in the SE task.

We further compared the enhanced prosodic FE results of the *multi-task* and baseline SE approaches with respect to three factors: gender, noise type, and SNR. The comparison results are listed in Table II. From the table, we first note that the intensity and f0 features exhibited very similar trends. When the gender factor was considered, the improvement in the *multi-task* SE system over the baseline SE system was clearer for female speakers. Next, when the noise types and SNR levels were considered, the *multi-task* SE system provided more improvements under challenging conditions, namely, low SNR and nonstationary noise types.

C. Qualitative Analysis of the Enhanced Features

In addition to the quantitative comparison of the *multi-task* and baseline SE systems, we also conducted a qualitative

study of the benefits of the unified feature enhancement approach. We selected one utterance from the test data and plotted the spectrogram, waveform, intensity, and f0 contours under three conditions: clean, enhanced by the baseline SE system, and enhanced by the *multi-task* SE system. The plots are shown in Figs. 2(a)–2(d), respectively.

From the spectrogram, waveform, and intensity plots shown in Fig. 2, identifying the differences between the *multi-task* and baseline SE systems is not easy. Because the intensity contour had a relatively high correlation with the waveform and spectrogram, the intensity constraint might have contained partial overlapping information with the spectral feature, which could explain why the improvement in the intensity estimation by the *multi-task* system presented in Section IV-B was not significant.

On the other hand, we can obviously see the distinction in the f0 contour extraction. The baseline SE system lost most f0 information and even failed to retain the ending portion of the utterance (see the green line in the f0 plot in Fig. 2(d)), thereby resulting in bad performance. This phenomenon indicated that using the spectral-based feature only could not effectively capture the f0 information. Introducing an additional f0 constraint during the training could resolve this problem, as shown by the red line in Fig. 2(d). These advantages and insights are obtained by evaluating different types of features, which cannot be obtained from the PESQ and STOI scores.

V. Conclusions and Future Work

In this study, we have investigated three learning approaches for a unified acoustic feature enhancement and found that the *multi-task* method achieves the best result. By imposing additional *praat* feature constraints, we are able to obtain improved intensity and f0 estimation from the enhanced speech without losing the STOI and PESQ scores. Speech-related application systems that rely on spectral and prosodic features, such as speech emotion recognition and gender recognition, may benefit from the unified feature enhancement framework.

Our immediate future work is to concatenate the back-end speech-processing techniques (e.g., emotion recognition, VAD, and VC) with the front-end unified feature enhancement framework to build end-to-end noise-robust speech application systems, and determine whether the unified feature enhancement can indeed improve the system performance under noisy environments. Moreover, we will keep on experimenting different acoustic features such as formant and MFCC to increase the generality of the unified feature enhancement.

Acknowledgment

This work was partly supported by MOST Taiwan Grants 108-2634-F-008-004 and 108-2634-F-001-004.

References

[1] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, 2015, pp. 91–99..

[2] 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, no. 4, pp. 745–777, 2014.

[3] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Robust automatic speech recognition: a bridge to practical applications." *Academic Press*, 2015.

[4] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[5] G. S. Bhat and C. K. Reddy, "Smartphone based real-time super Gaussian single microphone speech enhancement to improve intelligibility for hearing aid users using formant information," in *Proc. EMBC*, 2018, pp. 5503–5506.

[6] I. Panahi, N. Kehtarnavaz, and L. Thibodeau, "Smartphone-based noise adaptive speech enhancement for hearing aid applications," in *Proc. EMBC*, 2016, pp. 85–88.

[7] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[8] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2017.

[9] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop on Speech Coding*, 1999, pp. 165–167.

[10] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.

[11] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, 2017, pp. 2008–2012.

[12] M. Kolboek, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. IEEE Workshop on Spoken Language Technology*, 2016, pp. 305–311.

[13] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[14] S. Wang, K. Li, Z. Huang, S. Siniscalchi, and C.-H. Lee, "A transfer learning and progressive stacking approach to reducing deep model sizes with an application to speech enhancement," in *Proc. ICASSP*, 2017, pp. 5575–5579.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. Interspeech*, 2013, pp. 436–440.

[17] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, 2017, pp. 136–140.

[18] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. Interspeech*, 2015, pp. 3274–3278.

[19] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement." in *Proc. Interspeech*, 2016, pp. 3768–3772.

[20] S.-W. Fu, T.-W.Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[21] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 3364–3368.

[22] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "Highquality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.

[23] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[24] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proc. Speech Prosody*, 2016, pp. 84–88.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr´e, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[26] C.-H. Lin, M.-C. Wu, C.-L. You, C.-Y. Chiang, Y.-R. Wang, and S.-H. Chen, "Prosody modeling of spontaneous Mandarin speech and its application to automatic speech recognition," in *Proc. Speech Prosody*, 2016, pp. 1034–1037.

[27] S. Ghannay, Y. Est`eve, N. Camelin, C. Dutrey, F. Santiago, and M. Adda-Decker, "Combining continuous word representation and prosodic features for ASR error prediction," in *Proc. SLSP*, 2015, pp. 84–95.

[28] J. Orozco-Arroyave, F. H¨onig, J. Arias-Londo˜no, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. N¨oth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016.

[29] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multiobjective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. Interspeech*, 2015, pp. 1508–1512.

[30] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2158–2168, 2014.

[31] W. Li, N. Chen, M. Siniscalchi, and C.-H. Lee, "Improving mandarin tone mispronunciation detection for non-native learners with soft-target tone labels and blstm-based deep models," in *Proc. ICASSP*, 2018, pp. 6249–6253.

[32] H.-L. S. Wang, N. Y.-H. Wang, I.-C. Chen, and Y. Tsao, "Auditory identification of frequency-modulated sweeps and reading difficulties in chinese," *Research in developmental disabilities*, vol. 86, pp. 53–61, 2019.

[33] M. Huang, "Development of Taiwan Mandarin hearing in noise test," *Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[34] G. Hu, "100 nonspeech environmental sounds,[online] available:http://web.cse.ohiostate.edu/pnl/corpus/hunonspeech," *HuCorpus.html*, 2004.

[35] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[36] W.-C. Lin and C.-C. Lee, "Computational analyses of thin-sliced behavior segments in session-level affect perception," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2018.

[37] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P.* 862, 2001.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.