AUTHOR GUIDELINES FOR ICASSP 2024 PROCEEDINGS MANUSCRIPTS

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

Syllable stress is a fundamental aspect of speech that plays a crucial role in enhancing clarity, intelligibility, and meaning in communication. Accurate stress placement not only aids in distinguishing similar-sounding words but also contributes to the natural rhythm and emotional expression of speech, which are essential for effective human-computer interaction. In speech-based applications, maintaining these prosodic features is vital for achieving natural and seamless communication between humans and machines. However, many of these applications are trained on clean speech, limiting their effectiveness in real-world environments where noise is present. To address this challenge, speech enhancement techniques have been developed to remove noise from speech signals. Despite these advancements, it is critical to assess how well these methods preserve the original attributes of speech, particularly syllable stress. This study focuses on analyzing the impact of noise removal on stress patterns, with the aim of evaluating the effectiveness of speech enhancement methods in maintaining the prosodic features necessary for clear and natural human-computer interaction.

Index Terms— One, two, three, four, five

1. INTRODUCTION

Human-computer interaction (HCI) systems have become integral to everyday life, with speech serving as the primary mode of communication due to its accessibility, regardless of the user's literacy level. These systems rely heavily on the clarity and precision of spoken input to function effectively. In this context, spoken language cues such as syllable stress play a crucial role in conveying the intended message. Syllable stress, in particular, significantly influences lexical meaning, especially in stress-timed languages like English. While native English speakers naturally acquire this skill, non-native speakers often struggle with syllable stress due to the influence of their native language. To assist non-native speakers, Computer-Assisted Language Learning (CALL) systems have been developed to automatically detect stress patterns in syllables and provide guidance. For these systems to be effective, the automatic syllable stress detection module must be highly robust to ensure accurate guidance. However, since these modules are typically trained on clean, conditioned data, they can be prone to errors in noisy, real-world environments, which are often unavoidable.

To address the challenge of noise in real-world environments, two strategies can be employed: (1) training models with extensive, diverse noisy data to build robustness against noise, or (2) removing the noise from the speech before using for the downstream task. The first approach is often impractical due to its high computational cost, so in the literature, many systems prefer the second approach, known as speech enhancement. Speech enhancement (SE) is a process that improves speech quality and intelligibility by mitigating noise or reverberation. It is commonly used as a preprocessing step in applications such as automatic speech recognition (ASR), speech emotion recognition (SER), and automatic speaker verification (ASV) to enhance performance. As a preprocessing step, SE models have shown considerable improvements in certain speech applications. However, in some scenarios, the performance has been degraded compared to using noisy speech directly. This degradation in performance is often due to the loss of certain speech characteristics that are vital for the downstream tasks. To mitigate this issue, recent studies have explored the approach of jointly modeling the SE process with the downstream tasks, such as ASR, ASV, and SER. Integrating SE with these tasks aims to preserve essential speech features while improving the overall performance of the system.

Despite the extensive research on SE models in various speech applications, their impact on stress patterns has never been discussed. Specifically, it is unclear how noise affects these patterns and to what extent SE models can preserve or alter them. Furthermore, the literature lacks a thorough investigation into how SE models influence stress detection in noisy conditions, highlighting the need for further research in this area. Syllable stress is a key aspect in English, where every word has one syllable that is more prominent or relatively emphasized than any other syllable in the same word. This emphasized syllable is termed a stressed syllable. Automatic syllable stress detection involves classifying syllables into two categories (stressed/unstressed), making it a two-class classification problem. Several approaches for syllable stress detection have been proposed, primarily leveraging stress cues such as pitch, intensity, and duration [**?**], [**?**], [**?**]. Many machine learning (ML) models like boosting, bagging, decision trees, and support vector machines were used over the acoustic features computed on these stress cues for stress detection task [**?**]. An attention-based neural network combined with bidirectional LSTMs was used for stress detection, leveraging Mel frequency cepstral coefficients (MFCCs), energy, and pitch features [**?**]. Ruan et al. [**?**] utilized a transformer network for stress detection, while Mallela et al. [**?**] introduced a novel methodology with a joint optimization technique using variational autoencoders. All these studies have demonstrated the effectiveness of these methods under clean speech conditions. Despite these efforts, the challenge of accurately detecting stress patterns in noisy conditions and the effect of SE models in retaining stress cues is still not addressed.

The main objectives of this study are to: 1) examine the performance of stress detection under noisy conditions across different SNRs, 2) verify the impact of various state-of-the art SE models in preserving stress patterns amidst noise and their influence on the accuracy of stress detection, and 3) conduct a perceptual study to determine if the results of automatic stress detection align with perception-based outcomes.

Thanks to XYZ agency for funding.

2. DATASET

For this study, we utilize two datasets from the ISLE corpus: 1) German: This dataset includes 3,733 speech recordings from 23 non-native English speakers of German origin, and 2) Italian: This dataset comprises 3,981 speech recordings from 23 non-native English speakers of Italian origin.

Data Processing Phoneme alignments for both datasets are first computed using an automatic force alignment process. These alignments are then meticulously reviewed and corrected by a team of five linguists to ensure accuracy. Next, syllable alignments are derived from the phoneme alignments using the P2TK syllabification tool. The same group of linguists annotates syllable stress, ensuring that every word contains exactly one stressed syllable. Train and test sets are partitioned considering speaker nativity, age, gender, and language proficiency, to maintain a balanced distribution of stressed and unstressed syllables across both the sets [**?**]. We consider only polysyllabic words with two or more syllables for the experiments [**?**].

3. METHODOLOGY

In this section, we first describe the speech enhancement (SE) models considered in this study. Next, we outline the pipeline for the stress detection task that integrates these SE models. We then explain the procedure for the perceptual study, conducted separately to explore the relationship between the outcomes of the human perception study and the automatic stress detection task.

3.1. Speech Enhancement models (SE):

Speech enhancement is the task of maximizing the perceptual quality of speech signals, particularly by eliminating background noise [1]. Given the importance of SE models as a preprocessing step for real-time speech-based applications, numerous methods have been proposed in the literature to address this problem. Most early works are formulated in the time-frequency (spectrogram) domain and aim to approximate the clean spectrogram, from which they obtain the enhanced speech signal using the inverse short-time Fourier transform (iSTFT). However, the STFT is a generic transformation that may not be optimal for this task, and inaccuracies in phase reconstruction can limit the quality of the enhanced audio [2]. Additionally, these spectrogram-based approaches can be impractical for real-time and low-latency applications due to the need for high-resolution spectrograms, which require a long temporal window for STFT computation.

Adressing these limitations, several methods have been developed that learn representations directly from raw speech signals using deep learning [3, 4, 5]. While these approaches avoid some issues of spectrogram-based methods, they can still introduce unpleasant speech distortions and phonetic inaccuracies [6]. In contrast, diffusion models, a category of generative modeling, progressively denoise and refine the speech signal through iterative steps. This method enhances speech quality and naturalness while effectively reducing noise, making diffusion models particularly promising for challenging enhancement scenarios.

In this study, we consider three state-of-the-art speech enhancement models for noise removal in the automatic syllable stress detection task, considering factors such as latency, model complexity, speech quality, and the type of modeling (discriminative vs. generative). The details of each selected SE model are discussed in detail following this.

1) Dual-Signal Transformation LSTM Network (DTLN): The Dual-Signal Transformation LSTM Network (DTLN) is introduced in [7] for real-time speech enhancement (one frame in, one frame out). This model combines the short-time Fourier transform (STFT) with a learned analysis and synthesis basis in a stacked-network architecture, using fewer than one million parameters. The network proposed in DTLN enables robust extraction of information from magnitude spectra and incorporates phase information from the learned feature basis. It is trained on 500 h of noisy speech created on the part of the Librispeech corpus [8], and the noise signals originated from the Audioset corpus [9], Freesound, and DEMAND corpus [10] with SNR ranging from -5 to 25 dB.

2) DEMUCS-based real-time speech enhancement (Denoiser): DEMUCS is a novel architecture proposed for music source separation in [11]. To overcome the limitations of the Conv-Tasnet [2], which is the state-of-the-art approach for speech enhancement by noise removal, [1] proposed a real-time version of the DEMUCS architecture adapted for speech enhancement. It consists of a causal model, based on convolutions and LSTMs, with a frame size of 40ms, and a stride of 16ms. To enhance audio quality, the model operates directly on waveforms and utilizes hierarchical generation with U-Net-like [4] skip connections. The DEMUCS model was trained for 400 epochs on the Valentini dataset [12] and for 250 epochs on the DNS dataset [13].

3) Conditional diffusion probabilistic model (CDiffuSE): This work investigates diffusion probabilistic models [14], a class of generative models for speech enhancement. Diffusion probabilistic models convert clean input data to an isotropic Gaussian distribution in a step-by-step diffusion process and, in a reverse process, gradually restore the clean input by predicting and removing the noise introduced in each step of the diffusion process. These models, in their vanilla formulation, assume isotropic Gaussian noise in each step of the diffusion process as well as the reverse process. However, in realistic conditions, the noise characteristics are usually non-Gaussian, which violates the model assumption when directly combining the noisy speech signal in the sampling process. Addressing this, [6] proposed CDiffuSE, a conditional diffusion probabilistic model that can explore noise characteristics from the noisy input signal explicitly and thereby adapts better to non-Gaussian noise statistics in real-world speech enhancement problems. Also, it is shown that CDiffuSE maintains strong performance when regression-based approaches such as Demucs [1] and Conv-TasNet [2] collapse.

3.2. Automatic syllable stress detection (SE):

Figure 1 illustrates the entire pipeline for the stress detection task. We start with clean speech signals and generate four sets of noisy data with varying signal-to-noise ratios (SNRs): 0 dB, 5 dB, 10 dB, and 20 dB, using white Gaussian noise. Each set is processed individually for the experiments. From these noisy speech signals, we produce enhanced speech signals using three different SE models: CDiffuSE, Denoiser, and DTLN, as discussed earlier. Next, we perform syllable segmentation on each speech signal using syllable timestamps. For each syllable segment, we extract syllable-level features, which are then used to train a syllable stress detection classifier.

3.3. Perceptual study:

Syllable stress in English is crucial as it can change a word's grammatical category. While automatic stress detection helps identify

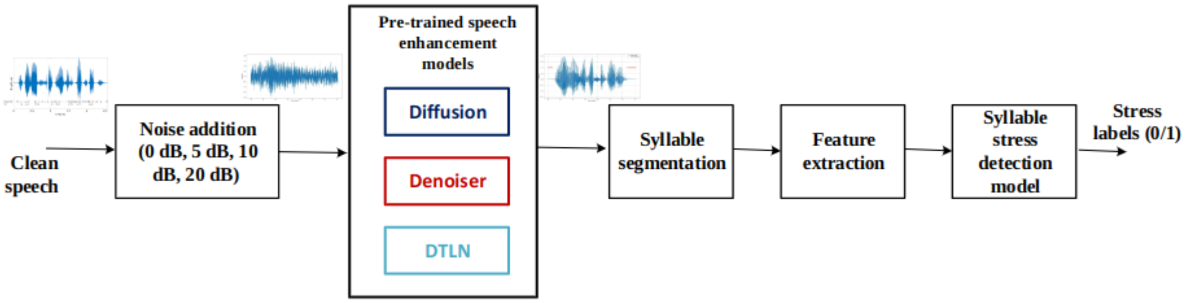

Clean speech
Noise addition (0 dB, 5 dB, 10 dB, 20 dB)
Pre-trained speech enhancement models
Diffusion
Denoiser
DTLN
Syllable segmentation
Feature extraction
Syllable stress detection model
Stress labels (0/1)

Fig. 1. Block diagram of the proposed approach for syllable stress detection

stress patterns, whether SE models effectively retain these patterns remains a key question. To complement the automatic detection process, we conduct a perceptual study to evaluate whether SE-enhanced audio retains the original stress pattern as perceived by human listeners. In this study, 25 participants aged 20-25 assess 50 audio samples, each representing a word that could belong to two grammatical categories depending on the stress placement. For each word, we provide a clean reference audio and three enhanced versions produced by different SE models applied to a 5 dB noisy signal of the same word. Participants are asked to select the audio most similar to the reference. This study aims to validate whether SE-enhanced audio retains the original stress pattern without introducing artifacts and to identify which SE model best preserves the stress pattern. Additionally, we aim to determine whether the performance of SE models in the automatic detection task aligns with human perception.

4. EXPERIMENTAL SETUP

In this section, we first provide an overview of the features used in our experiments: 1) heuristics-based acoustic and context features, and 2) self-supervised wav2vec 2.0 representations. We then discuss the architectural details of the classification models employed in the proposed approach.

4.1. Feature extraction:

Acoustic features (A) and Context features (C): Syllable stress typically relies on three prominence measures: intensity, duration, and pitch, centered around the most sonorous sound unit in the syllable. Acoustic features are derived from contours like short-time energy, representing these prominence measures. However, short-time energy contours can vary greatly across unstressed syllables and may introduce peaks due to surrounding sounds [**?**]. To address this, a sonority-based contour is proposed by combining sonority-motivated cues [**?**] with the short-time energy contour. Inspired by the work in [**?**], we compute 19-dimensional statistical features, referred to as acoustic features (A) in the following sections. In literature, it is stated that context significantly influences stress perception. In [**?**], a 19-dimensional binary context feature vector is proposed for each syllable, capturing information about the syllable nucleus type, the preceding and following phoneme categories, and the word's position relative to pauses in the sentence.

Self-supervised representations (wav2vec 2.0): Recently, self-supervised models like wav2vec 2.0 [**?**] and Hubert [**?**] have gained significant attention for learning speech representations without manual labeling. Wav2vec 2.0, in particular, is widely used for various speech tasks due to its ability to capture spectral, phonetic, speaker, and semantic features from raw speech data [**?**, **?**]. In this work, we utilize 768-dimensional wav2vec 2.0 features for our experiments.

Acoustic + Context	**Condition**	**SNR**			
		0 dB	**5 dB**	**10 dB**	**20 dB**
GER	**Noisy**	92.48	92.6	92.86	92.94
	Diffusion	**92.65**	**92.75**	**92.9**	**93.55**
	Denoiser	92.2	92.6	92.9	93.4
	DTLN	91.8	92.1	92.35	92.6
	Clean	93.36			
ITA	**Noisy**	92.3	92.58	92.86	92.46
	Diffusion	**92.4**	**92.65**	**92.89**	**93**
	Denoiser	90.3	90.8	91.1	91.45
	DTLN	90.45	90.95	91.15	91.3
	Clean	92.54			

4.2. Model architecture details:

In this study, we experiment with two different state-of-the-art classification models each belonging to different category following the work [**?**]. DNN with 5 layers [hidden units: 64, 32, 16, 4, 1] and VAE+DNN . The DNN excels in learning task-specific representations and capturing intricate, non-linear relationships within the features. While DNNs perform exceptionally well, they can benefit from enhanced robustness against data variations and complexities that naturally occur in real-world datasets [26]. To strengthen this, we integrate Variational Autoencoders (VAE) [27], which are highly effective in learning comprehensive data distributions and generating meaningful latent representations. VAEs, composed of an encoder and decoder, produce latent vectors that enhance the input features for DNNs. In this study, we train by jointly optimizing the loss functions of both VAE and DNN following the work in [**?**]. This combined model, referred to as VAE+DNN, ensures the learning of robust, task-specific representations, leading to improved performance. Binary cross-entropy is the loss function for all models.

5. RESULTS AND DISCUSSION

To achieve the best rendering both in printed proceedings and electronic proceedings, we strongly encourage you to use Times-Roman font. In addition, this will give the proceedings a more uniform look. Use a font that is no smaller than nine point type throughout the paper, including figure captions.

In nine point type font, capital letters are 2 mm high. **If you use the smallest point size, there should be no more than 3.2 lines/cm (8 lines/inch) vertically.** This is a minimum spacing; 2.75 lines/cm (7 lines/inch) will make the paper much more readable. Larger type

2*Acoustic + Context	2*Condition	SNR			
		0 dB	**5 dB**	**10 dB**	**20 dB**
5***GER**	**Noisy**	92.48 (89)	92.6 (89.22)	92.86 (89.78)	92.94 (90.02)
	Diffusion	**92.65** **(89.3)**	**92.75** **(89.5)**	**92.9** **(89.9)**	**93.55** **(90.5)**
	Denoiser	92.2 (88.8)	92.6 (89.25)	92.9 (89.5)	93.4 (89.95)
	DTLN	91.8 (88.2)	92.1 (88.7)	92.35 (89.3)	92.6 (89.7)
	Clean	93.36 (90.42)			
5***ITA**	**Noisy**	92.3 (88.92)	92.58 (89.22)	92.86 (89.24)	92.46 (89.4)
	Diffusion	**92.4** **(88.95)**	**92.65** **(89.25)**	**92.89** **(89.5)**	**93** **(89.55)**
	Denoiser	90.3 (85.75)	90.8 (86.35)	91.1 (87.2)	91.45 (87.5)
	DTLN	90.45 (85.8)	90.95 (86)	91.15 (86.8)	91.3 (87.1)
	Clean	92.54 (89.36)			

		SNR (dB)			
-2***Wav2vec 2.0**	-2***Condition**	**0 dB**	**5 dB**	**10 dB**	**20 dB**
	Noisy	87.12 (82.1)	91.52 (87.9)	93.72 (91.38)	94.34 (92.16)
	Diffusion	89.4 (85.2)	**93.2** **(90.26)**	**94.23** **(91.64)**	**94.8** **(92.9)**
	Denoiser	**91.42** **(87.46)**	92.9 (90.1)	94.1 (91.58)	94.3 (91.8)
	DTLN	89.18 (84.46)	92.5 (88.7)	93.5 (90.9)	94.1 (91.7)
-5***GER**	**Clean**	94.38 (91.62)			
	Noisy	85.96 (79.9)	91.4 (86.32)	93.3 (89.32)	94.28 (91.44)
	Diffusion	90.3 (86.8)	91.8 (88)	**93.4** **(89.5)**	**94.5** **(91.8)**
	Denoiser	**91.4** **(86.64)**	**92.74** **(88.76)**	92.9 (89.08)	94.22 (89.38)
	DTLN	86.48 (81.24)	91.34 (86.38)	92.6 (88.62)	93.72 (90.7)
-5***ITA**	**Clean**	93.63 (91.23)			

sizes require correspondingly larger vertical spacing. Please do not double-space your paper. TrueType or Postscript Type 1 fonts are preferred.

The first paragraph in each section should not be indented, but all the following paragraphs within the section should be indented as these paragraphs demonstrate.

6. MAJOR HEADINGS

Major headings, for example, ”1. Introduction”, should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (”.”) after the heading number, not a colon.

6.1. Subheadings

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

6.1.1. Sub-subheadings

Sub-subheadings, as in this paragraph, are discouraged. However, if you must use them, they should appear in lower case (initial word capitalized) and start at the left margin on a separate line, with paragraph text beginning on the following line. They should be in italics.

7. PRINTING YOUR PAPER

Print your properly formatted text on high-quality, 8.5 x 11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (12 mm) empty at the BOTTOM of the page and follow the top and left margins as specified. If the last page of your paper is only partially filled, arrange the columns so that they are evenly balanced if possible, rather than having one long column.

In LaTeX, to start a new column (but not a new page) and help balance the last-page column lengths, you can use the command “\pagebreak” as demonstrated on this page (see the LaTeX source below).

8. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

9. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Colors may be used, but they should be selected so as to be readable when printed on a black-only printer.

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeX document, below is an example of how to do this [15].

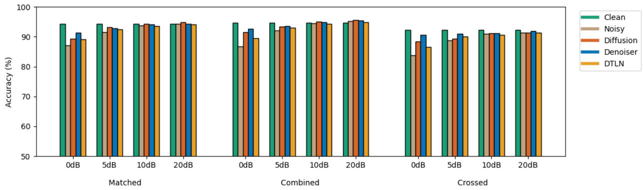


Fig. 2. *
Caption 1

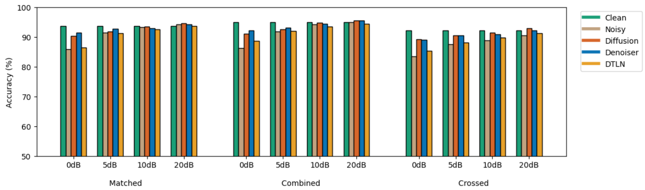


Fig. 3. *
Caption 2

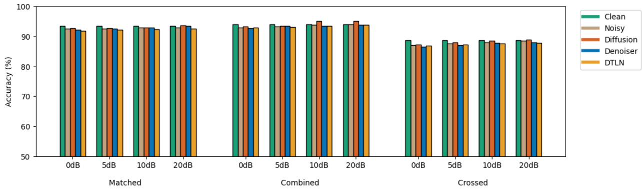


Fig. 4. *
Caption 3

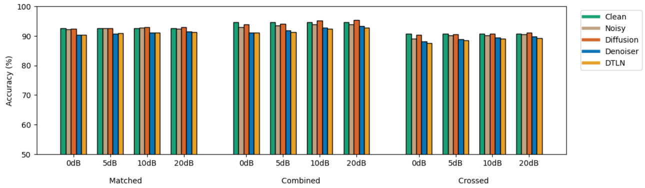


Fig. 5. *
Caption 4

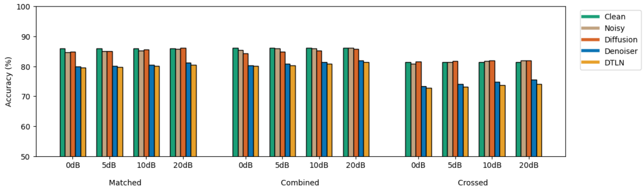


Fig. 6. *
Caption 5

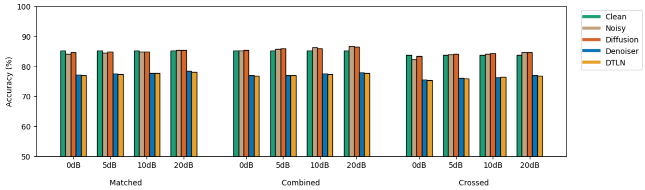
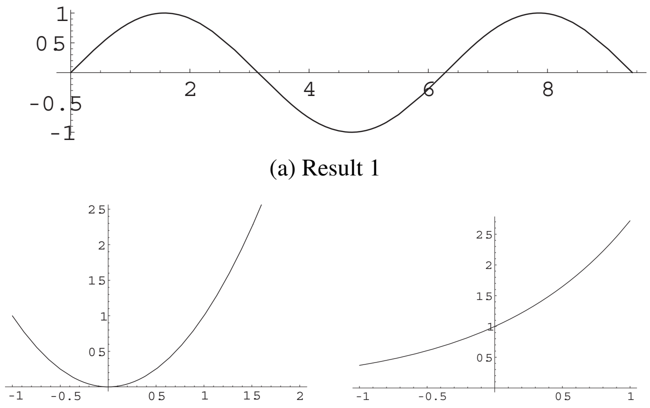


Fig. 7. *
Caption 6

Fig. 8. A 2x3 grid of images with captions.

10. FOOTNOTES

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).



(a) Result 1

(b) Results 3

(c) Result 4

Fig. 9. Example of placing a figure with experimental results.

11. COPYRIGHT FORMS

You must submit your fully completed, signed IEEE electronic copyright release form when you submit your paper. We **must** have this form before your paper can be published in the proceedings.

12. RELATION TO PRIOR WORK

The text of the paper should contain discussions on how the paper's contributions are related to prior work in the field. It is important to put new work in context, to give credit to foundational work, and to provide details associated with the previous work that have appeared in the literature. This discussion may be a separate, numbered section or it may appear elsewhere in the body of the manuscript, but it must be present.

You should differentiate what is new and how your work expands on or takes a different path from the prior studies. An example might read something to the effect: "The work presented here has focused on the formulation of the ABC algorithm, which takes advantage of non-uniform time-frequency domain analysis of data. The work by Smith and Cohen [15] considers only fixed time-domain analysis and the work by Jones et al [16] takes a different approach based on fixed frequency partitioning. While the present study is related to recent approaches in time-frequency analysis [3-5], it capitalizes on a new feature space, which was not considered in these earlier studies."

13. REFERENCES

List and number all bibliographical references at the end of the paper. The references can be numbered in alphabetic order or in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [16]. An additional final page (the fifth page, in most cases) is allowed, but must contain only references to the prior literature.

14. REFERENCES

[1] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[2] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] Yi Luo and Nima Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network.," in *Interspeech*, 2018, pp. 342–346.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[5] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[6] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7402–7406.

[7] Nils L Westhausen and Bernd T Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[10] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*. AIP Publishing, 2013, vol. 19.

[11] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[12] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[13] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[14] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[15] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article title," *Journal*, vol. 62, pp. 291–294, January 1920.

[16] C.D. Jones, A.B. Smith, and E.F. Roberts, "Article title," in *Proceedings Title*. IEEE, 2003, vol. II, pp. 803–806.