ROBUST AUDIO-VISUAL MANDARIN SPEECH RECOGNITION BASED ON ADAPTIVE DECISION FUSION AND TONE FEATURES

Hong Liu, Zhengyan Chen, Wei Shi

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
{hongliu, chenzhengyan, pkusw}@pku.edu.cn

ABSTRACT

Audio-visual speech recognition (AVSR) integrates both audio and visual information to perform automatic speech recognition (ASR), which improves the robustness of human-robot interaction systems especially in noise environments. However, few methods and applications have paid attention to AVSR in tonal languages, in which the linguistic feature can play an important role as well as visual information. In this work, we propose a method for AVSR in Mandarin based on adaptive decision fusion as well as making full use of tone features. Firstly, we introduce tone features calculated by Constant Q trasform (CQT) and put them into a CNN-based audio network together with Mel-Frequency Cepstral Coefficient (MFCC) audio features. Then, the visual features are extracted by Discrete Cosine Transform (DCT) from mouth regions in video frames and modeled by an LSTM-based visual network. Finally, an adaptive decision fusion network combines the outputs from both streams to make final predictions. Experimental results on the PKU-AV2 dataset show that the tone features can significantly improve the robustness of Mandarin speech recognition systems, and the adaptability of the proposed method to various noise environments.

Index Terms— Audio-visual speech recognition, tone feature extraction, decision fusion

1. INTRODUCTION

Automatic speech recognition (ASR) has been applied to a wide range of intelligent devices such as service robots, mobile phones, etc. Although most speech recognition systems perform well in clean conditions, it remains a challenge when applied to real-world environments with dramatically changing noise. Since visual information such as lip and tongue movements can play a complementary role for speech understanding, audio-visual speech recognition (AVSR) has shown noticable improvements over audio-only speech recognition in both clean and noise conditions [1–3].

In the past decades, significant research efforts have been made to perform AVSR [4–6]. Hu et al. proposed a recurrent temporal multimodal RBM to learn temporal joint representation efficiently [7] . Chung et al. applied an attention mechanism based on LSTM to fuse features from mouth ROIs and MFCCs [8] . More recently, an end-to-end audio-visual model is proposed to simultaneously extract features directly from mouth regions and raw waveforms [9].

Most existing studies on audio-visual fusion use feature-level fusion to model audio and visual information jointly. However, the combined feature will be dramatically affected by acoustic noises. Some works apply decision fusion by assigning fixed weights for each modality [10]. Wu et al. introduced a neural network to estimate weights from both outputs to achieve adaptive decision fusion under different noise environments [11]. On the other hand, few works have been proposed to focus on speech recognition in tonal languages such as Chinese, Vietnamese, Yorùbá, Igbo, etc [12–15], in which Mandarin Chinese is the most widely used language around the world. In tonal languages, words can differ in tones (like pitches in music), consonants and vowels. These linguistic features are less infected by noises, which will further improve the robustness of AVSR systems.

In this paper, an audio-visual Mandarin speech recognition model based on adaptive decision fusion and tone feature is proposed. In audio stream, Mel-Frequency Cepstral Coefficients (MFCC) and Constant Q Transform (CQT) features are concatenated as audio features. To the best of our knowledge, this is the first work that considers CQT as tone features for tonal language speech recognition. To get visual information, DCT features are extracted from lip regions in each frame. An audio CNN and a multi-layer LSTM are designed to model the audio and visual features, respectively. The audio CNN learns features from the time and frequency domains simultaneously through 2D convolution operation. Meanwhile, an LSTM-based visual network is utilised to learn sequential representation from visual feature sequences. Another fusion network combines the outputs from both streams to perform adaptive decision fusion. Experimental results on the PKU-AV2 dataset demonstrate the robustness of our tone features and the adaptability of the proposed fusion method. The results under various types and levels of noise verify that the audio-visual model significantly outperforms the audio-only models especially in presence of high levels of noise.

This work is supported by National Natural Science Foundation of China (NSFC No.61673030, U1613209).

978-1-7281-6395-6/20/$31.00 ©2020 IEEE

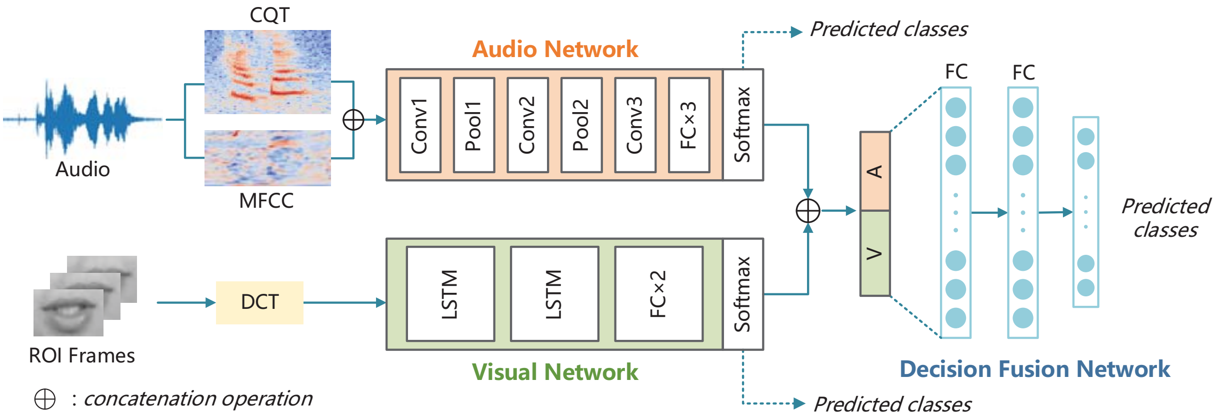


Fig. 1. Overview of the proposed audio-visual speech recognition architecture. The upper part is a CNN based audio stream, the lower part is an LSTM based visual steam, the right part is a decision fusion network.

2. FEATURE EXTRACTION

The input video is separated into audio waveforms and image frames, then features from each modality are extracted, respectively.

Audio Feature. Two audio features are extracted from the raw waveforms, i.e. MFCC and CQT features.

MFCC is a widely used audio feature in speech recognition, which is based on the human auditory perception properties. In our work, 40 dimensional MFCC is computed over a sliding window of 1024 samples and a frame shift of 512 samples, with the sample rate set to $22,050$Hz.

Motivated by music analysis tasks, we further use CQT to calculate a Constant Q spectrogram as tone features. The so-called Q factor reflects the selectivity of a filter used in time-frequency analysis and is defined as the ratio of its centre frequency and bandwidth:

$$Q = \frac{f_k}{f_{k+1} - f_k}, \tag{1}$$

where $k = 0, 1, ..., K$ is the frequency bin index and f_k is the centre frequency of bin k. When the bin frequencies are geometrically distributed as in the CQT transform. Denote b as the number of bins per octave, the formula (1) can be simplified to $Q = \left(2^{1/b} - 1\right)^{-1}$. The centre frequencies f_k can be calculated by:

$$f_k = 2^{k/b} f_0, \tag{2}$$

where f_0 is the fundamental frequency, and b determines the number of bins per octave. Compared to the short-time Fourier transform (STFT) which is used in MFCC, the CQT has a greater frequency resolution for lower frequencies but a greater temporal resolution for higher frequencies.

The CQT transform $X^{CQ}(k, n)$ of a discrete time-domain signal $x(n)$ is defined by:

$$X^{CQ}(k, n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j - n + N_k/2), \tag{3}$$

where $a_k(n)$ are basis functions including a window function (here uses the Hamming window), $*$ is the complex conjugate and N_k is a variable window length. More details of the CQT are presented in [16]. Note that different from music analysis tasks, we set $f_0 = 70$ Hz for CQT according to human voice fundamental frequency.

Visual Feature. For visual speech recognition (VSR), the region of interest (ROI) that contains most information is mainly the mouth region of the speaker. In our experiments, each frame in the video is converted to gray scale and face detection is performed using Viola-Jones algorithm [17]. The ROI of each video frame is extracted through 68 landmark points estimation [18], then resized to a fixed size of 45×30. Motivated by JPEG encoding, a two-dimensional separable DCT is applied to decompose the sinogram image into frequency coefficients. Then the top 55 coefficients with energy are selected in a Zig-zag pattern. Thus the visual features in a video $\mathbf{V}$ can be represented with a $55 \times T$ tensor, where T denotes the frame length of a video.

3. AUDIO-VISUAL MODELING

The architecture of the proposed audio-visual network is shown in Fig. 1. Our model is made up of three parts: a CNN-based audio network, a visual network based on LSTM, and a decision fusion network.

Audio Network. To learn spectral-temporal information simultaneously, we design an audio CNN containing three convolutional layers, two max-pooling layers and three fully connected (FC) layers. As shown in the upper part of Fig. 1, MFCC features and CQT features are concatenated together in temporal dimension, then the audio features are resized to 120×120 and put into the audio network. Note that the MFCC and CQT features are normalized separately before concatenation. The audio features are first put into a convolutional layer with kernel size of 3×3, then a max-pooling layer with kernel size of 2×2 is used to reduce the variability in the spectral-temporal domain. After that, the features are fed to another convolutional layer with kernel size of 5×5 followed by a max-pooling layer. Finally, after the third convolutional layer with kernel size of 5×5, three FC layers with the size of 1024, 512 and 100 followed by a softmax layer are used to get the audio probabilities $P(y_i^a|x^a)$.

Visual Network. In order to exploit sequential correlation, the visual features are fed to a two-layer LSTM which models the temporal dynamics of the input sequences. As shown in the lower part of Fig. 1, the visual network consists of two LSTM layers with 256 cells in each layer, followed by two FC layers with the size of 512 and 100 respectively. Note that LSTMs allow variable sequence lengths of input, thus the visual network is robust to the variation of speech velocity and word length. The predicted probabilities of the visual network can be denoted as $P(y_i^v|x^v)$.

Decision Fusion Network. As shown in the right part of Fig. 1, we integrate the audio and visual modalities based on single stream predictions, which is considered as late fusion or decision fusion. Compared to feature-level fusion, decision fusion does not require strictly synchronized inputs and is more robust to the variation of noise environments since audio and visual streams are modeled separately. Different from the weighting scheme that sums $P(y_i^a|x^a)$ and $P(y_i^v|x^v)$ by a weighting factor α to fuse both modalities [10], the output audio probabilities O_a and visual probabilities O_v from audio and visual network before softmax function are concatenated as a $1 \times 2C$ vector to be the input x^{av} of the decision fusion network, where $C = 100$ is the number of word classes. The decision fusion network contains 2 FC layers with the size of 256 and a softmax layer to generate audio-visual probabilities $P(y_i^{av}|x^{av})$. The audio, visual and the audio-visual fusion networks are trained with cross-entropy loss:

$$L_m = -\sum_{i=1}^{C} y_i \log P\left(y_i^m|x^m\right), \tag{4}$$

where $m \in \{a, v, av\}$ denotes audio, visual and audio-visual fusion network respectively.

4. EXPERIMENTS AND ANALYSIS

4.1. PKU-AV2 dataset

Since most datasets available are in English, we collected a novel audio-visual dataset by ourselves named PKU-AV2. The PKU-AV2 dataset contains 6000 utterances in total recorded by 6 subjects (3 males and 3 females), collected in an relatively noise-free environment with controlled normal light. Concerning the application on service robots, we choose 100 common foods and Chinese dishes as word corpus, such as "*huo guo*" (hotpot). To balance the length of the words, there are 35, 35, 30 words containing 2, 3 and 4 Chinese characters, respectively. Each word are repeated 10 times in Mandarin by each subject, in which 7 samples are randomly chosen for training. The dataset is recorded by a video camera at 30 frames per second with a resolution of 720×480 under the restriction that the mouth region is not occluded. The corresponding speech audio is synchronously recorded at a sampling frequency of 48 kHz with 16 bits per sample. Fig. 2 depicts some representation video frames in the PKU-AV2 dataset.

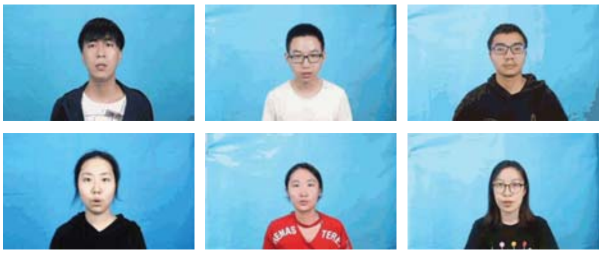


Fig. 2. Sample frames from the PKU-AV2 dataset.

4.2. Experimental Setting

In this work, the audio features are extracted by Librosa [19] Python package, all the networks are implemented in Keras on a NVIDIA GeForce GTX 1080 GPU. The stochastic gradient decent with a momentun of 0.9 is used to train audio and visual networks, while we use Adam optimizer for fusion network. The leaning rate and batch size are set to 0.001 and 64 respectively. The proposed model is evaluated under four types and five signal-to-noise ratios (SNRs) of noises. White Gaussian noise and speech babble noise are adopted from the Noisex92 database [20], another two kinds of noises are collected by ourselves, i.e., washing noise and music noise in order to apply our model in real environments such as a noisy restaurant. The four noise signals are added to the original audio signals with different SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB for the test data. We use clean audios to train the audio network and train the fusion network by adding white Gaussian noise with random SNRs between -10dB to 30dB.

4.3. Results and Analysis

The performance of audio-only system, visual-only system and the proposed audio-visual fusion method are evaluated for Mandarin speech recognition under four types of noises with 5 different SNRs (20 dB, 15 dB, 10 dB, 5 dB, 0 dB). As shown in Table 1, it is obvious that the visual-only system remains

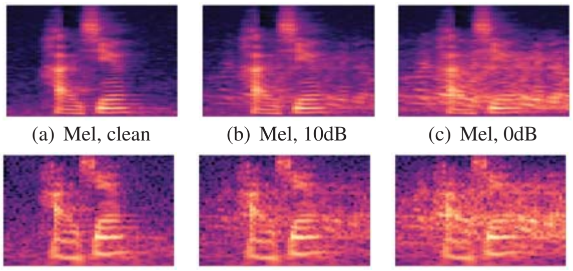
Table 1. Recognition accuracy (%) of the audio-only (A), visual-only (V) and audio-visual (A+V) models on PKU-AV2 dataset in different noise conditions.

	SNR(dB)	20	15	10	5	0
White noise	A (MFCC)	82.17	67.94	57.11	48.44	28.72
	A (CQT)	92.22	91.89	91.56	91.11	88.94
	A (MFCC+CQT)	94.06	93.94	93.61	92.88	89.50
	V	82.39	82.39	82.39	82.39	82.39
	A+V [10]	**95.72**	**95.67**	**95.5**	95.05	93.78
	A+V (ours)	95.67	**95.67**	95.44	**95.06**	**94.05**
Babble noise	A (MFCC)	79.39	67.67	56.67	48.44	27.68
	A (CQT)	91.83	91.38	90.72	87.38	70.56
	A (MFCC+CQT)	94.11	93.83	93.39	89.11	71.11
	V	82.39	82.39	82.39	82.39	82.39
	A+V [10]	95.72	95.72	95.72	93.61	89.16
	A+V (ours)	**95.78**	**95.83**	**96.00**	**93.89**	**89.78**
Washing noise	A (MFCC)	88.17	82.17	71.11	51.39	25.39
	A (CQT)	91.72	90.78	90.16	86.94	67.89
	A (MFCC+CQT)	93.05	93.05	92.17	89.06	71.39
	V	82.39	82.39	82.39	82.39	82.39
	A+V [10]	95.61	95.22	**95.00**	**93.28**	88.39
	A+V (ours)	**95.67**	**95.27**	94.61	93.11	**88.89**
Music noise	A (MFCC)	88.94	79.50	67.76	43.45	32.94
	A (CQT)	91.72	91.56	89.11	82.56	61.33
	A (MFCC+CQT)	93.72	92.61	89.94	84.39	62.50
	V	82.39	82.39	82.39	82.39	82.39
	A+V [10]	95.50	95.17	94.33	**91.83**	**87.77**
	A+V (ours)	**95.56**	**95.22**	**94.33**	91.61	87.73

constant performance (82.39%) for different noise intensities since visual information is unrelated to acoustic noises. We also run a GMM-HMM with same input visual features, the result (75.34%) shows our visual LSTM can model visual information more effectively. However, the performance of audio-only system is acceptable under low levels of noise (15 to 20dB), but decreases substantially with stronger noises. We also observe that the audio stream is most affected by music noise while less effected by white noise, that is because the CQT is closely related to the tone of the utterances, which is similar to pitch in music.

In order to investigate the adaptability of the proposed method to various noises, the decision fusion network is compared with linear weighting method used in [10]. Note that the α in [10] is set by traversing in range $[0, 1]$ with a step of 0.01 on the training data to choose the optimum weight for fusion in every noise environments, representing the best result in an ideal state of known noise. The Results from all types and SNRs of noises shown in Table 1 demonstrate that the audio-visual fusion outperforms both single modalities, especially under high levels of noise ($\leq$10dB), and the proposed decision fusion network outperforms the ideal results of weighting scheme in most noise conditions.

We also conducted experiments to evaluate the robustness and effectiveness of the MFCC features, CQT features and their combination. Fig. 3 shows some visualization of Mel-spectrogram and CQT-spectrogram of an audio sample under clean conditions and babble noise. The Mel-spectrogram becomes illegible with the presence of strong noise while the audio features in CQT-spectrogram remain relatively clear as in the lower part of Fig. 3(f). As shown in Fig. 4, the CQT and MFCC features have similar performance in noise-free conditions. However, the recognition accuracy of MFCC features decreases significantly in the presence of high levels of babble noise, while CQT remains more robust (up to 41.84% better than MFCC under 0dB). Similar results can be observed under other types of noises according to Table 1.


(a) Mel, clean (b) Mel, 10dB (c) Mel, 0dB

(d) CQT, clean (e) CQT, 10dB (f) CQT, 0dB

Fig. 3. Visualization of Mel-spectrogram and CQT-spectrogram under noise-free, 10dB and 0dB of babble noise.

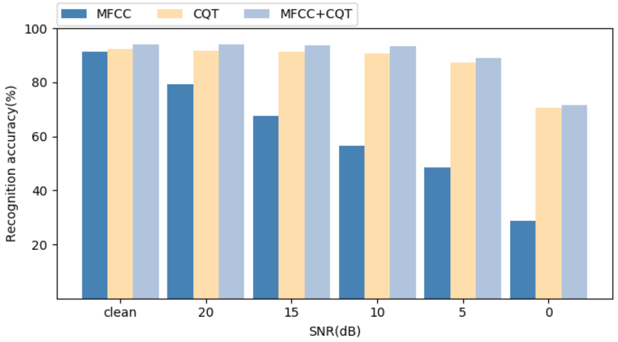


Fig. 4. Comparisons of audio-only speech recognition accuracy of MFCC features, CQT features and the combined features in clean conditions and various levels of babble noise.

5. CONCLUSIONS

In order to make full use of visual information and language features for speech recognition, a robust audio-visual Mandarin speech recognition method based on adaptive decision fusion and tone features is proposed. The audio network effectively models the time-frequency information from robust tone features based on Constant Q Transform. The visual network models the sequential information from mouth ROIs by a two-layer LSTM network. An adaptive decision fusion network integrates the audio and visual streams, making a significant improvement to single-modality speech recognition. Experiments on the PKU-AV2 dataset demonstrate the robustness and adaptability of our method for Mandarin AVSR in various noise environments. Our future work will extend the proposed method to AVSR in Vietnamese, Igbo and other tonal languages.

6. REFERENCES

[1] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE transactions on multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[2] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] Stavros Petridis and Maja Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2015.

[4] Weijiang Feng, Naiyang Guan, Yuan Li, Xiang Zhang, and Zhigang Luo, "Audio visual speech recognition with multimodal recurrent neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 681–688.

[5] Yuki Takashima, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki, Nobuyuki Mitani, Kiyohiro Omori, and Kaoru Nakazono, "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss.," in *Interspeech*, 2016, pp. 277–281.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[7] Di Hu, Xuelong Li, et al., "Temporal multimodal learning in audiovisual speech recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[8] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[9] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[10] Runwei Ding, Cheng Pang, and Hong Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4138–4142.

[11] Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, and Xuewu Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[12] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[13] Ngoc Thang Vu and Tanja Schultz, "Vietnamese large vocabulary continuous speech recognition," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 333–338.

[14] Hank Chang-Han Huang and Frank Seide, "Pitch tracking and tone features for mandarin speech recognition," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2000, vol. 3, pp. 1523–1526.

[15] Wei Zou, Dongwei Jiang, Shuaijiang Zhao, Guilin Yang, and Xiangang Li, "Comparable study of modeling units for end-to-end mandarin speech recognition," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 369–373.

[16] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[17] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[18] Vahid Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[20] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.