Hindawi
Mobile Information Systems
Volume 2022, Article ID 5116280, 6 pages
https://doi.org/10.1155/2022/5116280

Research Article

Application of the Artificial Intelligence Algorithm in the Automatic Segmentation of Mandarin Dialect Accent

Yufang Lai

Jiangxi Health Vocational College, Nanchang Medical College, Nanchang 330000, Jiangxi, China

Correspondence should be addressed to Yufang Lai; 2016120223@jou.edu.cn

Received 29 December 2021; Revised 27 January 2022; Accepted 1 February 2022; Published 24 February 2022

Academic Editor: Hasan Ali Khattak

In recent years, as the research objects of phonetics have expanded to accent and colloquial natural speech, the construction of the dialect accent Mandarin voice database has become another important research direction in the field of computer technology. Among them, voice segmentation is a time-consuming and laborious link in the construction of the voice database. The application of artificial intelligence technology helps to improve the construction efficiency of the Mandarin dialect voice database. Based on this, this article mainly researches the application of the artificial intelligence algorithm in the automatic segmentation of dialect accent Mandarin. This paper constructs a voice corpus of dialect accents and Mandarin Chinese and specifically describes the construction process of the voice corpus. This paper uses artificial intelligence algorithms, combined with the HMM (hidden Markov model), and Viterbi algorithm to propose a new method of automatic speech segmentation. This paper studies the automatic speech segmentation model, extracts the general parameters of the training data in the Mandarin corpus, and conducts HMM training. This paper conducts tests based on the voice of the test set to verify the accuracy of the method proposed in this paper. The experimental results show that, in the speech data of 60 people, the error range of each sentence time period is less than 5 ms accounting for 79.16%, less than 10 ms accounting for 82.96%, less than 20 ms accounting for 83.14%, and less than 50 ms accounting for 86.92%. It can be seen that the algorithm proposed in this paper can meet practical applications in automatic speech segmentation.

1. Introduction

Artificial intelligence technology has the ability of autonomous calculation, which can accurately extract feature parameters and identify and filter out accurate voice feature data. The error is small, and the accuracy is high. The extracted and analyzed voice features are more reliable and more useable and powerful. It can clearly be seen that the continuous development and maturity of artificial intelligence have promoted the deepening of the research field of speech recognition; especially, the research of automatic speech segmentation algorithm has received a lot of attention [1, 2]. With the help of computer technology and input-related model algorithms, the system can automatically divide the speech signal into a series of sound segments and classify them into specific acoustic units [3, 4]. Our country has a vast territory, and dialects between different places will inevitably affect the pronunciation of standard Mandarin, and the accent of Mandarin, which is between standard Mandarin and dialects, is produced from this [5]. At present, it is still very difficult to automatically segment Mandarin with dialect accent [6]. Therefore, to improve the accuracy of automatic segmentation of dialect accent Mandarin, it is very important to study the use of artificial intelligence algorithms for automatic speech segmentation.

Regarding the research of speech recognition and automatic segmentation, many scholars have carried out multiangle investigations. For example, Tom et al. improved the robustness of the acoustic speech measurement system through the study of automatic presegmentation algorithms of speech [7]; Savchenko studied the algorithm of guaranteeing the level of significance in the automatic segmentation of speech signals [8]; Hanani and Naser used the *X*-vector algorithm to conduct speech recognition research on spoken

Arabic dialects [9]. It can be seen that the research on speech recognition algorithms is relatively rich, but the research on automatic speech segmentation is still lacking. Therefore, this article studies the application of artificial intelligence algorithms in the automatic segmentation of dialect accent Mandarin speech, which is conducive to enriching theories about this field.

This paper takes the automatic speech segmentation algorithm as the research object, combined with artificial intelligence, and studies a new method of automatic segmentation of dialect accent Mandarin. This article first designs a dialect accent Mandarin speech corpus and introduces the construction process of the speech corpus. Then, this article uses artificial intelligence algorithm, combined with the HMM (hidden Markov model), fused with Viterbi algorithm, and proposes a new method of automatic speech segmentation. Finally, this paper tests the accuracy of automatic speech segmentation, which verifies the accuracy of the proposed method of dialect accent Mandarin speech segmentation.

2. Application of the Artificial Intelligence Algorithm in the Automatic Segmentation of Dialect Accent and Mandarin Speech

2.1. Design of the Dialect Accent Mandarin Speech Corpus

2.1.1. Speech Corpus. According to the speech corpus construction process, the corresponding information products must be stored in each process. The main steps for establishing the dialect accent Mandarin speech corpus are shown in Table 1.

2.1.2. Dialect Accent Mandarin Voice Sample Recording. After selecting the text corpus, a total of 221 samples of the text corpus were selected, and the text corpus needed to be recorded. Voice recording is a heavy task and a heavy workload, which requires the cooperation of multiple people. First of all, it needs to find a suitable pronunciation partner, be familiar with the content of the corpus, record the voice according to the requirements (environment, recording specifications, etc.), check and process repeatedly, and organize storage reasonably.

2.1.3. Corpus Collection. Voice collection is based on mobile phone recording, including major smart phone brands. Install the voice recorder app on the mobile phone. The recording environment includes indoor quiet environment and outdoor noise environment. The recording specifications are as follows:

(1) Quiet environment: the windows are closed, and there are no other noises such as human voices and TV

(2) Noise environment: various noises such as human voice, TV, air conditioning, and incoming windows

(3) Read in accordance with natural language flow, with fluent and clear pronunciation and without emotion, and pay attention to the pauses between each word, phrase, and sentence

(4) Channel: mono

(5) Depth: 16 bits

(6) Sampling rate: 8,000 Hz; format: WindowsPCM (.wav)

The basic requirement for recorders is to understand the content of the recording before recording and to be familiar with words, phrases, and sentences in advance. When recording, keep a distance of about 10 cm between the recorder and the mobile phone, and try to keep calm. After recording, the voice collected by each recorder will be put in a separate directory, named after the number of the speaker's data. For the convenience of collection, each person will speed up and slow down the commonly used dialect accents, Mandarin Chinese, and daily expressions. In normal speaking rate pronunciation recording, there is a pause between each word and sentence, which is convenient for subsequent nuclear pronunciation and sorting. Each recorded voice file is named with "corpus type + speaking rate," such as common text (fast).

2.1.4. Sorting and Nuclear Sound. For large-vocabulary speech recognition systems, considering the lack of speech data, context-dependent triphones are often used as modeling units in acoustic model modeling. However, the number of triphone units formed is large, so the demand for training speech data is also high, subsequently increased. For the original corpus collected for the first time, it must be used after many nuclear sounds and finishing processing [12]. Since each piece of corpus is a continuous recording file of a certain type of corpus, in order to ensure the correctness of the dialect accent in Mandarin and the subsequent smooth segmentation of the sentence corpus, after each recording, it is necessary to monitor the original voice file of the collected corpus. The collected data need to be separately detected to prevent the incorrect collection of dialect accent and Mandarin pronunciation or the occurrence of mistakes and more. If you find related problems, you need to make up or delete the problematic voice immediately.

2.1.5. Speech Preprocessing and Feature Extraction. In this process, the input sound is first divided into frames, and the sound is divided into many small segments so that there is a certain overlap between frames. Then, feature extraction is performed on the conversion of small segments, and a frame of waveform is turned into a multidimensional vector to achieve the effect of dimensionality reduction. What is used here is Mel-frequency cepstral coefficient (MFCC) for feature extraction from dialect accent Mandarin speech data. The overall process includes voice signal preprocessing, DFT/FFT processing, filter processing, logarithmic operation, and, finally, DCT cepstrum. In feature extraction, it is necessary to ensure that the feature parameters are representative and that each parameter is independent of each other.

Table 1: Information storage table generated by the speech corpus construction process.

Construction process	Information type	Information product	Storage form
Text corpus selection	Text	Text corpus information	Database
Voice recording	Text	Pronunciation partner information Original collected voice information	Database
	Voice	Voice information after segmentation processing Raw voice	File system
		Phonetic corpus	
Voice annotation	Text	Voice-corresponding labeling text information	Database

At the same time, it is necessary to use the HMM to convert the acoustic characteristics of the dialect Mandarin speech into the dialect Mandarin pronunciation consonants and finals. The model at this stage is called the acoustic model. After the sequence of the actual pronunciation consonants and vowels is obtained, decoding techniques such as language models can be used to convert the actual pronunciation consonants and vowel sequences into texts.

The decoding process calculates the acoustic model and language model scores for the given pronunciation consonant and final sequence and several hypothetical word sequences and uses the sequence with the highest overall output score as the recognition result.

2.1.6. Corpus Augmentation. The dialect accent Mandarin speech perturbs the speech rate, which changes the speech rate. The essence is to linearly stretch or compare the speech signal in the time domain. When the length of the speech signal changes, its framing situation will also be changed, and the shape of the disturbed speech signal in the frequency domain will also change to a certain extent. If the disturbed speech rate is slower than the original, its speech energy will shift to low frequencies; if the disturbed speech rate is faster than the original, its speech energy will shift to high frequencies. And because of the logarithmic relationship between the Mel frequency and the voice signal frequency, corresponding changes will also occur.

In the case of changes in the time domain and frequency domain, the characteristics extracted from the dialect accent Mandarin speech signal will be somewhat different from the original. Therefore, the speech after the speech rate disturbance can be used as new data to some extent. The original voice data can be expanded after mixing in the original data. In the process of corpus augmentation, the audio processing tool SoX is used to perturb the speed of speech batches. Once the noise is mixed into the speech data, the waveform and frequency spectrum are changed, so as to achieve data augmentation, and at the same time, it can also enhance the noise robustness of the dialect accent Mandarin speech database. This paper compares and analyzes the recognition results of the accent-related and accent-related acoustic models, studies the effectiveness of the accent-related decision tree clustering method in capturing accent characteristics, and uses it for the subsequent comparative analysis experiments of the accent-related acoustic models based on artificial intelligence.

2.2. Automatic Voice Segmentation

2.2.1. Automatic Segmentation Process. HTK is a toolkit for building HMM, which mainly has the functions of data preparation, model training, and optimization evaluation. Therefore, this research uses the HTK toolkit to complete the automatic segmentation of dialect accent Mandarin speech. First, extract the general parameters, fundamental frequency, and duration of the training data in the dialect accent Mandarin speech corpus, and perform HMM training. HMM is defined as follows.

Assume that N is the length of time the sample observes, and $X = (X1, X2, \ldots, XN)$, $Y = (Y1, Y2, \ldots, YN)$, $x = (x1, x2, \ldots, xN)$, and $y = (y1, y2, \ldots, yN)$, $i_n \in \Phi$, $y_n \in V$, $1 \le n \le X$; the joint distribution of X and Y satisfies the hidden Markov condition, as shown in the following:

$$P(Y = y, X = x) = \pi_{i_1 y_1} b_{i1} a_{i_2 y_2} \ldots b_{i_{N-1} y_{N-1}} b_{i_N y_N}. \quad (1)$$

Among them, a_{ij} represents the further transition probability of X_N, and b_{ij} is the probability that the observation sequence takes the value v_j at time n when the state process is taken at time n. π is the initial state probability vector, as shown in formula (2).

Use Viterbi segmentation algorithm to initialize and recurse the input model and observation model, so as to find the optimal path, get the most likely category of each frame of dialect accent Mandarin audio data, and realize automatic segmentation, and finally, the segmentation results are automatically classified.

3. Experimental Research on the Automatic Segmentation Algorithm of Dialect Accent and Mandarin

3.1. Experimental Tools. This experiment used Baidu AI Studio deep learning platform, Kaldi open-source speech recognition system, and PyTorch software.

3.2. Experimental Program

3.2.1. Experiment 1: Dialect Accent Mandarin Speech Recognition. The corpus dataset is divided into ten mutually exclusive subsets, each with 1200 sentences of the speech corpus, and the speech data of the test set and the training set are divided in a ratio of 1 : 9. In the experiment, the cross-validation of each piece of data was repeated ten times in turn, and the obtained accuracy rates were averaged.

Table 2: Dialect speech recognition results.

Model	Number of test sets	Correct quantity	Number of errors	Accuracy (%)
This article	1200	1174	26	97.83
CNN-HMM	1200	1008	192	84
FSMN	1200	1080	120	90

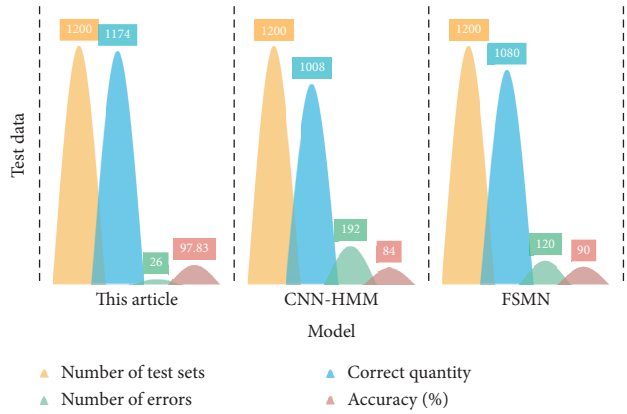

Figure 1: Dialect speech recognition results.

3.2.2. Automatic Voice Segmentation. Use 10, 30, and 60 dialect accent Mandarin speech data as the training set, which are 100, 300, and 600 speech data, respectively. In this paper, the maximum accuracy of automatic segmentation of Mandarin accent based on HMM Viterbi segmentation algorithm model, Bayesian information criterion model (BIC), and likelihood criterion (maximum likehood) is compared.

3.3. Evaluation Criteria. The evaluation standard for the accuracy of automatic speech segmentation is the segmentation accuracy, and its calculation method is shown in the following formula:

$$W.\text{Acc} = 1 - \text{WER} = 1 - \left(\frac{S}{T}\right) \times 100\%. \quad (2)$$

Among them, W.Acc represents the correct rate of voice segmentation, WER represents the error rate of voice segmentation, S represents the number of errors in the segmentation of dialect accents in Mandarin, and T represents the number of samples.

4. Experimental Results of the Automatic Speech Segmentation Algorithm

4.1. Dialect Speech Recognition Results. Use the same test set to test on the acoustic model based on the hidden Markov model based on the neural network (CNN-HMM) and the feedforward sequential memory network (FSMN) model. The cross-validation was repeated ten times, the accuracy of the three-algorithm dialect speech recognition was compared, and the average was taken. The results are shown in Table 2.

It can be seen from Figure 1 that this paper correctly identified 1,174 dialect voices, and the model achieved a recognition rate of 97.83% on the cross-validation test set.

Table 3: The accuracy of BIC voice automatic segmentation (%).

Data	≤5 ms	≤10 ms	≤20 ms	≤50 ms
10	53.21	61.27	73.16	81.29
30	51.20	57.69	70.47	80.31
60	49.78	45.74	68.24	76.58

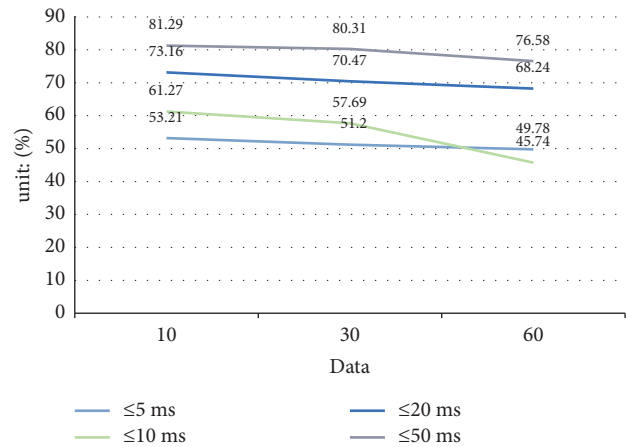

Figure 2: The accuracy of BIC voice automatic segmentation (%).

The recognition rate on the CNN-HMM is 84%, and the recognition rate on the FSMN model is 90%. On the same dataset, the model constructed in this paper can achieve a 13% lower error rate than the CNN-HMM acoustic model and a 7% lower error rate than the FSMN model, which significantly improves the performance of the model.

4.2. BIC's Automatic Voice Segmentation Accuracy. BIC converts the task of finding change points into the task of selecting models. The experimental results of automatic speech segmentation are shown in Table 3.

It can be seen from Figure 2 that, in the speech data of 10 people, the error range of each sentence time period is less than 5 ms, accounting for 53.21%, less than 10 ms accounting for 61.27%, less than 20 ms accounting for 73.16%, and less than or equal to 50 ms accounting for 81.29%; in 30 people's voice data, the error range of each sentence time period was 51.20% for 5 ms, 57.69% for 10 ms, and 70.47% for 20 ms, which was less than 50 ms accounting for 80.31%; among 30 people's voice data, the error range of each sentence time period was 49.78% for 5 ms, 45.74% for 10 ms, and 68.24% for 20 ms and less than 20 ms. 50 ms accounted for 76.58%.

4.3. ML's Automatic Voice Segmentation Accuracy. Similarly, the results of automatic segmentation of ML's speech are sorted, and the results are shown in Table 4. In the

Table 4: ML's automatic voice segmentation accuracy (%).

Data	≤5 ms	≤10 ms	≤20 ms	≤50 ms
10	56.23	67.24	78.28	87.16
30	54.45	58.34	70.53	85.24
60	52.18	51.68	69.46	76.29

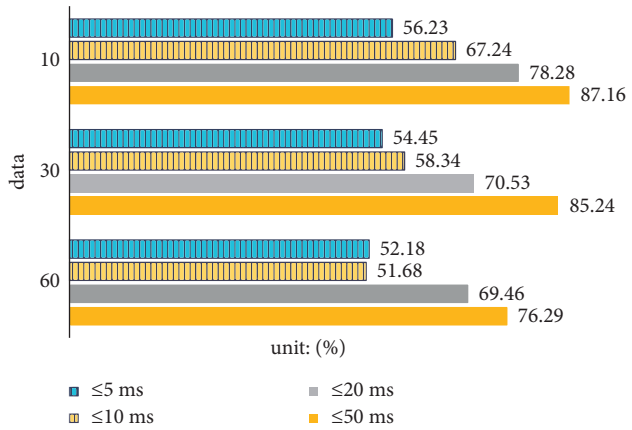

Figure 3: ML's automatic voice segmentation accuracy (%).

Table 5: The accuracy of the HMM combined with Viterbi segmentation algorithm for automatic speech segmentation (%).

Data	≤5 ms	≤10 ms	≤20 ms	≤50 ms
10	82.26	83.79	85.72	89.46
30	79.49	80.45	83.91	87.73
60	79.16	82.96	83.14	86.92

ML model, in the speech data of 10 people, the error range of each sentence time period is less than 5 ms, accounting for 56.23%. 67.24% are less than 10 ms, 78.28% are less than 20 ms, and 87.16% are less than 50 ms; in 30 people's voice data, the error range of each sentence period is less than 5 ms, accounting for 54.45%. 58.34% are less than 10 ms, 70.53% are less than 20 ms, and 85.24% are less than 50 ms; in 60 people's voice data, the error range of each sentence period is less than 5 ms, accounting for 52.18%. The proportion of less than 10 ms is 51.68%, the proportion of less than 20 ms is 69.46%, and the proportion of less than 50 ms is 76.29%.

It can be seen from the data in Figure 3 that, in the voice data of 10, 30, and 60 people, the accuracy of automatic segmentation of the voice of the dialect accent of ML is higher than that of the BIC. In 10 people's voice data, the accuracy rate of ≤50 ms is improved by about 5%.

4.4. Accuracy of Automatic Voice Segmentation Based on the HMM Combined with the Viterbi Segmentation Algorithm. In this experiment, the same speech data are used for model training, and the results are calculated, as shown in Table 5. In the speech data of 10 people, the error range of each sentence period is 82.26% less than 5 ms, 83.79% less than 10 ms, 85.72% less than 20 ms, and 89.46% less than 50 ms. In the speech data of 30 people, the error range of each sentence cycle is less than 5 ms (79.49%), less than 10 ms (80.45%), less than 20 ms (83.91%), and less than 50 ms (87.73%). In the speech data of 60 people, the error range of less than 5 ms accounts for 79.16%, that of less than 10 ms accounts for 82.96%, that of less than 20 ms accounts for 83.14%, and that of less than 50 ms accounts for 86.92% in each sentence period.

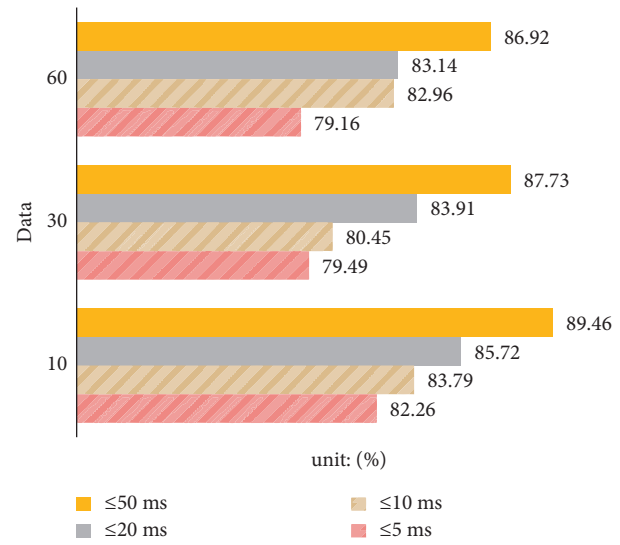

Figure 4: The accuracy of the HMM combined with Viterbi segmentation algorithm for automatic speech segmentation (%).

It can be seen from the data in Figure 4 that the segmentation accuracy of the HMM combined with Viterbi segmentation algorithm is significantly higher than that of the other two models. This shows that the segmentation algorithm proposed in this paper can achieve the automatic segmentation of dialect accent Mandarin.

5. Conclusions

Artificial intelligence broadens the theoretical research and application range of audio segmentation technology. Through research, this paper has completed the following tasks: introduces the construction process of the dialect accent Mandarin speech corpus and designs the corpus; this paper collects dialect accent Mandarin speech data from the processes of voice sample recording, corpus collection, collation, and nuclear pronunciation. This paper combines the HMM model and Viterbi segmentation algorithm to construct an automatic segmentation model for dialect accent Mandarin speech based on artificial intelligence algorithms; based on the experimental data, this paper compares and analyzes the automatic segmentation accuracy of HM-Viterbi segmentation algorithm model, BIC model, and ML model and proves the high efficiency of the HM-Viterbi segmentation algorithm model [10, 11].

Data Availability

The data underlying the results presented in this study are available within the manuscript.

Disclosure

The author confirms that the content of the manuscript has not been published or submitted for publication elsewhere.

Conflicts of Interest

The author declares no conflicts of interest.

Authors' Contributions

The author saw the manuscript and approved it to submit to the journal.

References

[1] O. Zealouk, H. Satori, N. Laaidi, M. Hamidi, and K. Satori, "Noise effect on Amazigh digits in speech recognition system," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 885–892, 2020.

[2] A. Masmoudi, F. Bougares, M. Ellouze, and E. Yannick, "Automatic speech recognition system for Tunisian dialect," *Language Resources and Evaluation*, vol. 52, no. 2, pp. 1–19, 2017.

[3] W. Ying, L. Zhang, and H. Deng, "Sichuan dialect speech recognition with deep LSTM network," *Frontiers of Computer Science in China: English Edition*, vol. 14, no. 2, p. 10, 2020.

[4] S. Shivaprasad and M. Sadanandam, "Dialect recognition from Telugu speech utterances using spectral and prosodic features," *International Journal of Speech Technology*, vol. 5, pp. 1–10, 2021.

[5] G. Agarwal and H. Om, "Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 1–32, 2021.

[6] M. Gomathy, "Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 155–163, 2021.

[7] B. Tom, L. Laura, A. Paavo, and V. Erkki, "Automatic presegmentation of running speech improves the robustness of several acoustic voice measures," *Logopedics Phoniatrics Vocology*, vol. 28, no. 3, pp. 101–8, 2003.

[8] V. V. Savchenko and A. V. Savchenko, "Guaranteed significance level criterion in automatic speech signal segmentation," *Journal of Communications Technology and Electronics*, vol. 65, no. 11, pp. 1311–1317, 2020.

[9] A. Hanani and R. Naser, "Spoken Arabic dialect recognition using X-vectors," *Natural Language Engineering*, vol. 26, no. 6, pp. 1–10, 2020.

[10] V. Bhardwaj and V. Kukreja, "Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions," *Applied Acoustics*, vol. 177, no. 3, Article ID 107918, 2021.

[11] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Transactions on Multimedia*, vol. 99, p. 1, 2020.

[12] M. Hamidi, H. Satori, O. Zealouk, and S. Khalid, "Amazigh digits through interactive speech recognition system in noisy environment," *International Journal of Speech Technology*, vol. 23, no. 2, pp. 1–9, 2020.