IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 26, 2020, accepted September 6, 2020, date of publication September 15, 2020,
date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024218

WaveNet With Cross-Attention for Audiovisual Speech Recognition

HUI WANG, FEI GAO, YUE ZHAO, (Member, IEEE), AND LICHENG WU
School of Information Engineering, Minzu University of China, Beijing 100081, China

Corresponding authors: Hui Wang (wanghui_610919@163.com) and Yue Zhao (zhaoyueso@muc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976236, and in part by the MUC 111 Project.

ABSTRACT In this paper, the WaveNet with cross-attention is proposed for Audio-Visual Automatic Speech Recognition (AV-ASR) to address multimodal feature fusion and frame alignment problems between two data streams. WaveNet is usually used for speech generation and speech recognition, however, in this paper, we extent it to audiovisual speech recognition, and the cross-attention mechanism is introduced into different places of WaveNet for feature fusion. The proposed cross-attention mechanism tries to explore the correlated frames of visual feature to the acoustic feature frame. The experimental results show that the WaveNet with cross-attention can reduce the Tibetan single syllable error about 4.5% and English word error about 39.8% relative to the audio-only speech recognition, and reduce Tibetan single syllable error about 35.1% and English word error about 21.6% relative to the conventional feature concatenation method for AV-ASR.

INDEX TERMS Cross-attention mechanism, multimodal speech recognition, WaveNet model, end-to-end-model.

I. INTRODUCTION

In our daily life, man-machine interaction interface for all kinds of devices is necessary. Speech is the most natural and convenient interaction mode between human and machine. In particular in a natural and noisy setting, audiovisual speech recognition is most effective way by using lip movement information to complement acoustic speech to recognize speech content [1]. However, multimodal speech recognition remains challenging for how to combine two data streams with different frame rates in feature space.

WaveNet is a deep generative model with very large receptive fields. It is composed of dilated causal convolutional layers, which enlarges the receptive field by skipping input values with a certain step. It is powerful for modelling the long-term dependency on speech data. It has been efficiently applied for speech generation, text-to-speech, and speech recognition [2].

In this work, we extent the WaveNet to multimodal speech recognition. To capture the effective fusion feature and address the alignment of two data streams in different frame rates, we introduce the cross-attention mechanism to WaveNet, and combine the Connectionist Temporal Classification (CTC) with WaveNet for end-to-end multimodal speech recognition. For AV-ASR, the early fusion and middle fusion between two modal data are widely used. The conventional method is to concatenate the acoustic feature with visual feature after making the number of visual frames equal with the audio frames. It is prone to fuse acoustic features with non-correlated visual features, and cannot effectively align the frames between two modes. In this work, cross-attention mechanism is explored to place at input layer or hidden layer in WaveNet, to automatically learn the weights of visual frames near the current audio frame. Several weighted visual feature frames form a visual context vector, which is concatenated with the current acoustic feature frame. The visual feature frames with large score provide more effective information for acoustic features, and they match with the current audio frame much more. Owing to the large receptive field of WaveNet model in high layers, it is more difficult to align the visual hidden feature frames with acoustic hidden feature frames. So we try to introduce cross-attention into high layer of WaveNet for middle fusion.

The proposed model was evaluated on Tibetan and English audiovisual speech data respectively. Compared to the traditional feature concatenation and audio-only speech recognition based on WaveNet, our model achieved better performance.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

Our work has three contributions: (i) we introduced the cross-attention mechanism to align two modals data in feature space. (ii) we explored the effects on cross-attention mechanism for early fusion and middle fusion in WaveNet. (iii) we explore the video frame shift for cross-attention calculation to improve the speech recognition performance and computation speed.

II. RELATED WORK

There is a lot of literature for complementing the visual information, like lip motion patterns, to acoustic speech to improve the performance of automatic speech recognition. Many research works have proposed the effective methods to extract powerful visual features for lip-reading, including active shape models (ASMs) [3], active appearance models (AAMs) [4], discrete cosine transform (DCT), principal component analysis (PCA) and discrete wavelet transform (DWT) [5], [6], etc. ASMs locate and track lip contours which use a priori knowledge about shape deformation from the statistics of a training set which was labelled by hand. AAMs contain a statistical model of the shape and grey-level appearance of the object of interest which can generalize to almost any valid example. ASMs and AAMs both contain priori information, and they extract human face information from pictures. PCA, DCT and DWT are data analysis and digital signal processing field methods which do not contain prior knowledge. The image features obtained by PCA, DCT and DWT are the characteristics of the entire picture.

Neural networks, such as deep neural (DNN), recurrent neural network (RNN) and long short-term memory (LSTM) have been introduced to the field of speech recognition [7]–[9] and AV-ASR [1], [10]–[16]. End-to-end neural networks are challenging the dominance of HMM as a core technology. Because end-to-end ASR has more advantages than DNN/HMM systems, especially for low-resource languages, it avoids the need for linguistic resources like dictionaries and phonetic knowledge [17], so more and more speech recognition research works are deployed with end-to-end method. The main approaches use the end-to-end to train encoder-decoder, which automatically learn acoustic/visual features directly from raw input data [1], [10]–[16] instead of using the manually defined features extracted by priori knowledge. Although these methods can learn all the required information directly from the data, raw images and speech signal will lead to high dimensional input. The work of [18] used a bottleneck layer to reduce the dimension of visual features, and the work of [19] used hand-crafted visual features of lip contour for the use of CNN-based feature, which demonstrated that lip landmark's movements are very effective visual features for when the size of the training dataset is limited. In this work, we also used the geometric distances of lip landmarks as visual features for visual information describing lip movements.

One of key points in AV-ASR is about the fusion of audio and visual information. Turbo decoder (TD) framework which is inspired by iteratively exchanging some kind of soft information between the audio and video decoders until convergence is applied in [20] to solve this problem in decision fusion way. The work of [16] focuses on fused information which introduces the gating layer to remove noisy or uninformative visual feature to solve the problem that the recognition accuracy is decreased for AV-ASR when speech is clean. Our method differs by introducing cross-attention mechanism between two modalities to reduce the effect of uncorrelated visual feature frame to acoustic feature frame.

The work of [1] proposed modality attention mechanism in the watch, listen, attend and spell (WLAS) model to replace the concatenation of two context vectors obtained by attending over individual modality to predict the output units. Not only that individual modality context vector is computed by attention mechanism, it also needs to compute the modality attention at each decoding step. The work of [15] correlates every frame of acoustic feature with visual context feature acquired by cross modality attention of two encoders over all video frames. These works have high computation cost. In our work, we use a sliding window to capture the related visual feature frames, which improves the computation speed compared with the calculation of the attention coefficient of all frames.

III. DATA

Our model is trained on two datasets: the first is an open and free Tibetan multi-dialect speech data set TIBMD@MUC [21], which contains 1803 Tibetan short videos. We randomly select 154 short videos as test data. The data is about daily conversation in Lhasa-Tibetan dialect. Acquired with a laptop equipped with a webcam and microphone, the speech data are corrupted with low environmental noise from classrooms and dormitories. All text corpora include a total of 1361 Tibetan syllables. These data sets can be downloaded from https://pan.baidu.com/s/1apXJgR53tNKZQ2LPCjDjqQ.

The second data set is TCD-TIMIT [22] which consists of high-quality audio and video footage of 62 speakers reading a total of 6913 phonetically rich sentences. 1730 and 180 sentences were selected as our training data and test data respectively.

All audio files are separated from videos by using the ffmpeg library. The audio files are converted to 16KHz sampling rate, 16bit quantization accuracy, and WAV format. 39 MFCC features of each observation frame were extract from speech data using a 128ms window with 96ms overlaps.

All video files have the frame rate 25 fps (frames per second). To keep the number of visual frame and audio feature frame from the same video consistent, we insert one visual frame every three visual frames by computing the average of the adjacent two frames. We used Dlib library to detect faces and then extract the 12 points (No.48-No.59 point as shown in Fig. 1) near the mouth. From the 12 mouth points, we compute the distance between left and right mouth corner as the width of mouth opening and 5 distances between 5 points on the upper lip and 5 points on the lower lip as the

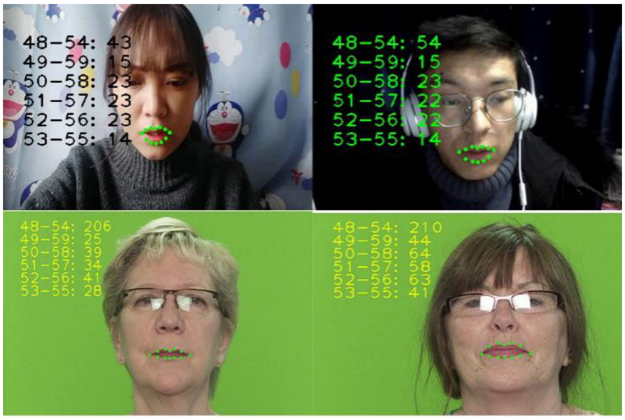


FIGURE 1. Visual features frames.

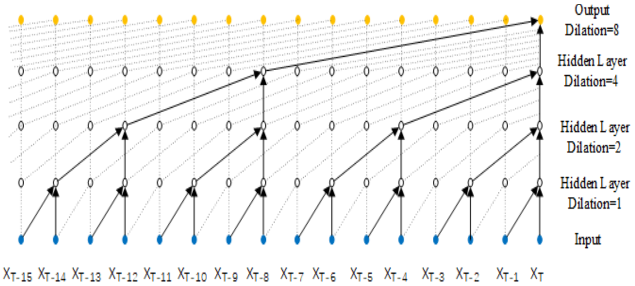


FIGURE 2. Dilated causal convolutional layers [2].

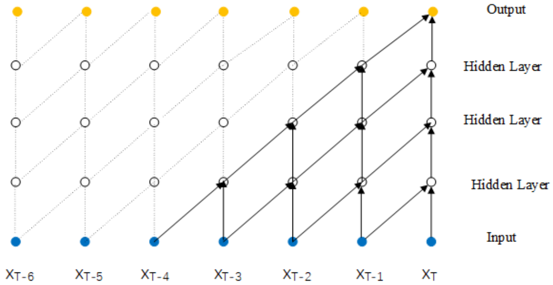


FIGURE 3. Standard causal convolutional layers [2].

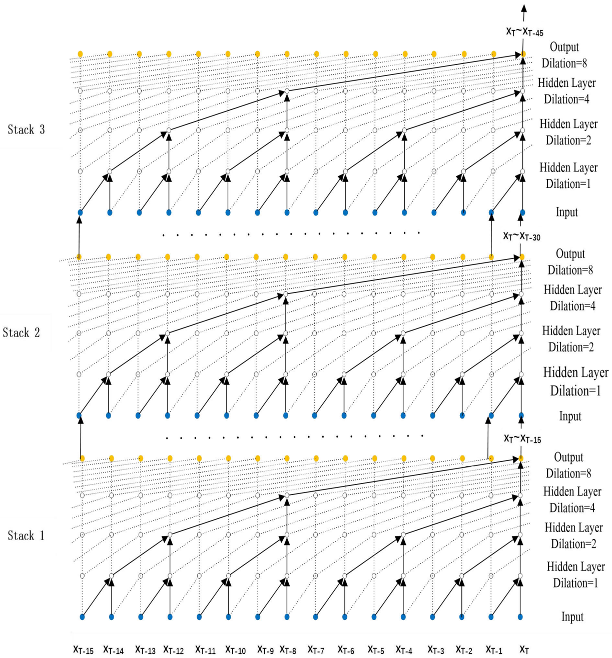


FIGURE 4. Stacked dilated causal convolutional layers. $x_m \sim x_n$ on arrows represents the receptive filed range of output nodes.

heights of mouth opening which is shown in Fig. 1. These six distances are taken as the visual feature for input to the models.

IV. METHOD

A. WaveNet-CTC

1) WaveNet

The WaveNet model is composed of stacked dilated causal convolutional layers [2]. The network models the joint probability of a waveform as a product of conditional probabilities defined as in equation (1).

$$p(x) = \prod\nolimits_{t=T-RF+1}^{T} p(x_t | x_{T-RF+1}, \cdots, x_{t-1}) \tag{1}$$

where *RF* represents the number of receptive filed of stacked dilated causal convolutional layers.

A stack of dilated causal convolutional layers with dilation $\{1, 2, 4, 8\}$ and filter length 2 is shown in Fig. 2. It is more efficient than standard causal convolution layers in Fig. 3 for increasing the receptive field, since the filter is applied over an area larger than its length by skipping input values with a certain step. The receptive field of a block of dilated causal convolutional layers is calculated by equation (2).

$$\begin{aligned} Receptive\_field_{block} = \sum\nolimits_{i=1}^{n} [(&Filter_{length} - 1) \\ &\times Dilation\_rate_i] + 1 \end{aligned} \tag{2}$$

where $Dilation\_rate_i$ refers to the dilation rate of i-th layer.

Stacking a few blocks of dilated causal convolutional layers creates a very large receptive field size. For example, 3 blocks of dilated convolution with the dilation $\{1, 2, 4, 8\}$ are stacked in Fig.4, where a $\{1, 2, 4, 8\}$ block has receptive field of size 16, and then the dilation repeats as $\{1, 2, 4, 8, 1, 2, 4, 8, 1, 2, 4, 8\}$. So, the stacked dilated convolutions have a receptive field of size 46. However, the standard causal convolution with 12 layers and filter width 2, the receptive field is only 13($=$ #layers $+$ filter length -1). The receptive field of the stacks of dilated convolutions is computed by equation (3).

$$Receptive\ field_{stacks} = S \times Receptive\ field_{block} - S + 1 \tag{3}$$

where S refers to the number of stacks.

WaveNet uses the same gated activation unit as the gated PixelCNN [23]. Activation function is given by equation (4).

$$h_t = tanh(W_{f,t} * x_t) \odot \sigma(W_{g,i} * x_i) \tag{4}$$

where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, and $\sigma(\cdot)$ is a sigmoid function. i is the layer index. f and g denote the filter and gate, respectively, and W is learnable weight. Also, WaveNet uses residual and parameterized skip connections to speed up convergence and enable training of much deeper models. More details on WaveNet can be found in [2].

2) CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist Temporal Classification (CTC) is an algorithm that trains a deep neural network for sequence labeling tasks, such as speech recognition and handwriting recognition, which allows the network to make label predictions at any point in the input sequence. Since it does not require the alignment between input sequence and target sequence, it removes the need for the pre-segmented data. Moreover, CTC directly outputs the probabilities of the complete label sequences, so the procedure of external post-processing is not required.

For speech recognition, CTC needs a softmax layer as the output layer of a deep neural network to produce posterior probabilities of frame-level labels for each label a_t in target set C which consists all output labels plus a ''blank'' symbol. Denote $\mathbf{z} = (a_1, a_2 \ldots\ldots a_T)$ as a frame-level label sequence (referred to as CTC path), $\mathbf{y}$ as transcript (a true target label sequence), $B^{-1}(\mathbf{y})$ as the preimage of $\mathbf{y}$ mapping all possible CTC paths $\mathbf{z}$ that result in $\mathbf{y}$. CTC loss function is defined as the negative log of sum of the probabilities of all possible CTC paths $\mathbf{z}$ that result in $\mathbf{y}$, as equation (5).

$$Loss_{CTC} = -\log p(\mathbf{y}|\mathbf{x}) = -\log \sum\nolimits_{\mathbf{z}\in B^{-1}(\mathbf{y})} p(\mathbf{z}|\mathbf{x}) \qquad (5)$$

where $\mathbf{x}$ is input sequence of speech features, and $p\,(\mathbf{z}\,|\,\mathbf{x})$ is the posterior probability of a frame-level label sequence $\mathbf{z}$ given by the product of the individual frame posteriors as equation (6).

$$p\,(\mathbf{z}\,|\,\mathbf{x}) = \prod\nolimits_{t=1}^{T} p(a_t|\mathbf{x}) \qquad (6)$$

Owing to exponentially growing the number of CTC paths with the frame length and the size of target label set , CTC loss uses the forward-backward algorithm to accelerate the process of mapping speech to a text sequence [24]. The forward-backward algorithm replaces the sum over all possible CTC paths with an iterative sum over paths corresponding to prefixes of a labelling $\mathbf{y}$. The recursive forward and backward variables are used for efficient computation in the iteration.

The decoding process of CTC in this paper is based on a beam search algorithm [25], which is more accurate than greedy search.

3) BASELINE

We adopt the architecture of WaveNet-CTC [26], [27] shown in Fig. 5 as the baseline model for audio-only speech recognition and audiovisual speech recognition with the conventional feature concatenation in our experiments. The original WaveNet in reference [2] was used for speech recognition directly on raw audio at sample level. However, considering the alignment of two modal data in different frame rates for audio-visual speech recognition, we made WaveNet operate at frame level. In Fig. 5 the model integrates WaveNet [2] with CTC loss to train WaveNet for end-to-end speech recognition.

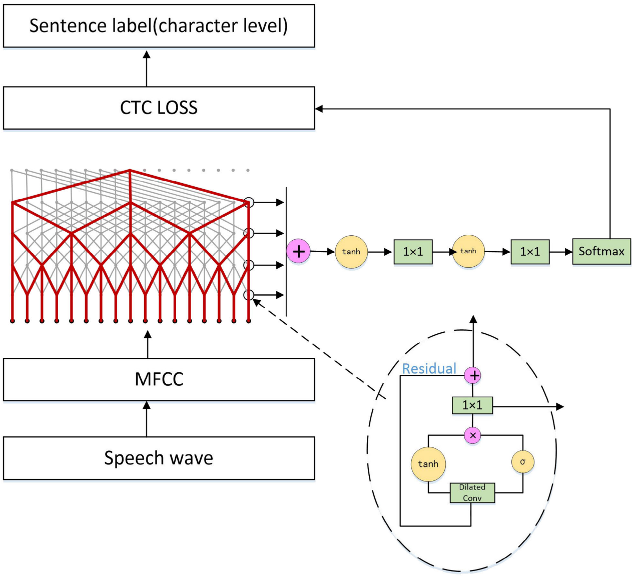


FIGURE 5. The architecture of WaveNet-CTC [26], [27].

WaveNet [2] consists of a softmax to compute the posterior probabilities of each speech frame, and then CTC loss is calculated based on the output of softmax to tune the parameters of WaveNet.

ང་ལ་སྒོར་བརྒྱད་ཡོད།

FIGURE 6. A Tibetan sentence (It means I have eight bucks).

Tibetan character is composed of single syllable. One or several syllables form Tibetan words. A Tibetan sentence is shown in Fig. 6, where syllables is separated by delimiter ''.'' and the sign ''|'' is used as the end sign of a Tibetan sentence. Each syllable are written in Tibetan alphabet from left to right, but it is different from English, in some syllables there is a vertical superposition in addition [21], which forms a two-dimensional planar character as shown in Fig. 7. A Tibetan syllable usually contains several consonant letters and a vowel sign corresponding to 7 parts, i.e., prescript, superscript, subscript, root, vowel sign, postscript and post-postscript.

Since Tibetan character is a two-dimensional planar character, Tibetan letters are not suitable directly as the modeling unit for the end-to-end model like English letters. Tibetan letters sequence is output one by one from model, which cannot form a two-dimensional Tibetan character. If end-to-end model are trained with Tibetan letters as modelling unit, the post-processing is need to form Tibetan character. So, in this work we used single syllable as the CTC modeling unit for Tibetan. However, English word is written with English letters from left to right, which is a one-dimensional character. We use English letters as modelling unit for end-to-end learning.

B. CROSS-ATTENTION MECHANISM

To automatically align the correlated visual feature frames to acoustic feature frames h_t, cross-attention layer is proposed

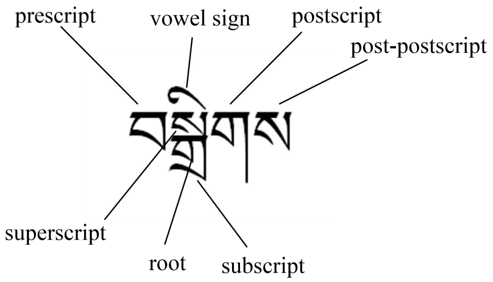


FIGURE 7. The structure of a Tibetan syllable.

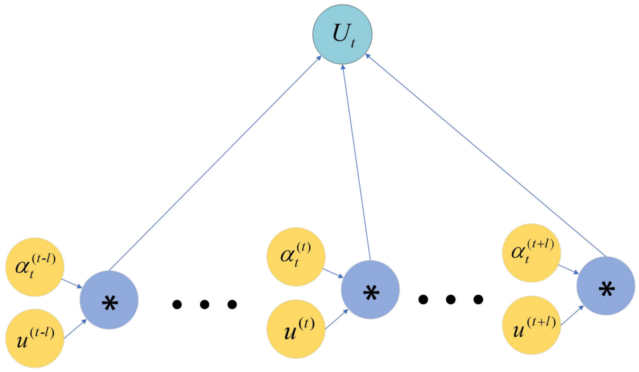


FIGURE 8. The schematic diagram of visual context vector calculation.

to create visual context vector U_t to fuse with the current acoustic frame h_t. The cross-attention layer produces a fused feature at each frame t by setting an ‘‘attention range’’ before and after the current video frame t, and producing output of visual context vector, and then concatenating it with current acoustic feature frame. Visual context vector captures the correlated visual frames for the current acoustic frame for the sake of alignment between two data streams.

The formula for calculating the visual context vector is shown in equation (7), and the schematic diagram is as Fig. 8.

$$U_t = \sum\nolimits_{s=t-l}^{t+l} \alpha_t^s \cdot u^s \tag{7}$$

where α_t^s is the attention weight, subject to $\alpha \geq 0$ and $\sum_s \alpha_t^s = 1$ through softmax normalization. l represents the length of sliding window before and after the current frame t, and u^s is visual feature frame. The α_t^s calculation method is defined as equation (8).

$$\alpha_t^s = \frac{exp(Score(h_t, u^s))}{\sum_{s=t-l}^{t+l} exp(Score(h_t, u^s))} \tag{8}$$

It represents the correlation of acoustic-visuals frame pair (h_t, u^s). Since automatic speech recognition depends on the particular context, the current audio frame just is related to several visual frames before and after it, the local attention is used to operate on a sliding window of $2l+1$ visual frames including the current visual frame u^t. $Score(\cdot)$ is computed as equation (9) by the MLP which is jointly trained with all the other component in end-to-end network [28]. Those u^s that have large score have more weights in visual context vector U_t, which are more correlated with the current acoustic frame h_t for alignment.

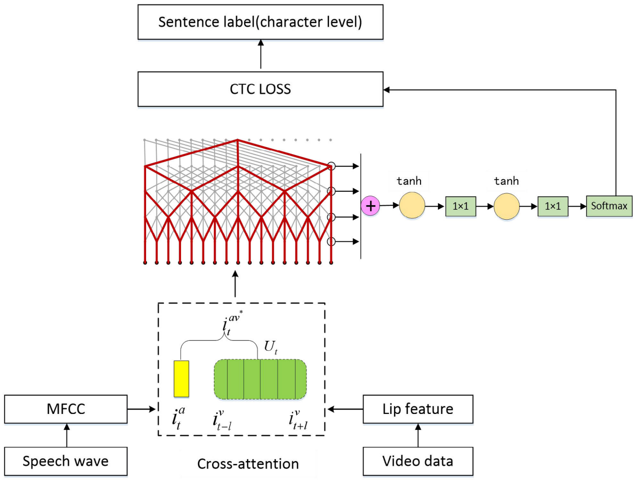


FIGURE 9. The architecture of WaveNet-CTC with cross-attention for input-layer fusion.

$$Score\left(h_t, u^s\right) = v_a^T \tanh(W_a[h_t; u^s]) \tag{9}$$

Finally, acoustic feature h_t is concatenated with U_t as the fused feature h_t^{av*}. Given the sliding window contains the visual frames after the current time, the fused features are fed only into the input layer before dilated causal convolutional layers as shown in Fig. 9, or output layer (softmax layer) after dilated causal convolutional layers in WaveNet as shown in Fig. 10. At input layer, the visual context feature was fused with MFCCs. At the high-layer, acoustic hidden feature was fused with high-layer visual context vector.

V. EXPERIMENT

We evaluated the Tibetan single syllable error rate and English word error rate of WaveNet-CTCs with cross-attention for Tibetan and English audio-visual speech recognition, and compared them with audio-only WaveNet-CTC (A-WaveNet-CTC), visual-only WaveNet-CTC (V-WaveNet-CTC), audio-visual WaveNet-CTC with feature fusion at input layer without attention (AV-WaveNet-CTC-I) and audio-visual WaveNet-CTC with feature fusion at high layer without attention (AV-WaveNet-CTC-H). The models of WaveNet-CTC with cross-attention are corresponding to two fusion places, which are feature fusion at input layer of WaveNet (AV-WaveNet-CTC-A-I) and the one at high-layer of WaveNet (AV-WaveNet-CTC-A-H).

A. EXPERIMENTAL SETTINGS

The WaveNet network consists of 15 layers, which are grouped into 3 dilated residual block stacks of 5 layers, and in each layer, original input was added with the output of a residual block and taken as new input into the next residual block to enhance the data abstraction of different depth levels of the network. In every stack, the dilation rate increases by a factor of 2 in every layer, starting with rate 1 (no dilation) and reaching the maximum dilation of 16 in the last layer. The filter size of causal dilated convolutions is 7. The number

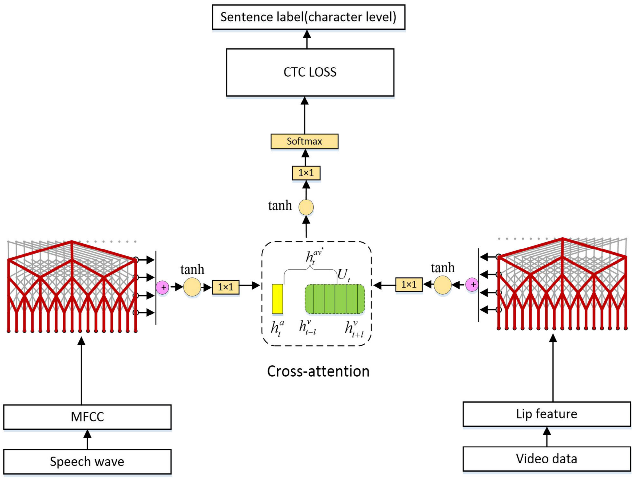


FIGURE 10. The architecture of WaveNet-CTC with cross-attention for high-layer fusion.

TABLE 1. Tibetan single syllable error rates of models.

Models	Single syllable error rate (%)
A-WaveNet-CTC	47.2
V-WaveNet-CTC	88.1
AV-WaveNet-CTC-I	77.8
AV-WaveNet-CTC-A-I-3	42.9
AV-WaveNet-CTC-A-I-5	46.0
AV-WaveNet-CTC-A-I-7	**42.7**
AV-WaveNet-CTC-A-I-10	53.9
AV-WaveNet-CTC-H	58.5
AV-WaveNet-CTC-A-H-3	73.6
AV-WaveNet-CTC-A-H-5	70.9
AV-WaveNet-CTC-A-H-7	75.4
AV-WaveNet-CTC-A-H-10	78.6

of hidden units in the gating layers is 128. The learning rate is 2×10^{-4}. The number of hidden units in the residual connection is 128.

In this work, the half size l of sliding window is from 3 to 10 frames. The sliding window with smallest size would contain original visual frames since visual frames are linearly interpolated in every 3 frames.

B. EXPERIMENTAL RESULTS

In training, we directly optimize the CTC loss between ground-truth syllables and predicted syllables sequence via the Adam optimizer. In evaluation, we measure the Levenshtein edit distance to calculate Tibetan single syllable error rate and English word error rate. The experimental results for Tibetan are shown in Table 1 and the results for English are shown in Table 2.

In Table 1 and Table 2, we can see that the error rates of visual-only speech recognition are greatly higher than the one of audio-only speech recognition, which indicate that the effective information carried by the video is very rare for speech recognition. The error rate of AV-WaveNet-CTC-I model for Tibetan is 77.8%, which is higher than A-WaveNet-CTC, but for English the error rate of AV-WaveNet-CTC-I is 39.4% which is lower than A-WaveNet-CTC, i.e. audio-only models. We analyze the reason why the results are different for Tibetan and English, and we think it maybe is that the Tibetan video quality is not standard and good as English data, since the head pose or facial expressions are various in Tibetan video, which influence the width and heights measure for mouth opening. The visual information disrupts acoustic information and reduce the recognition rate for Tibetan. However, for English data, the visual information incorporating into the audio information improves the speech recognition accuracy.

TABLE 2. English word error rates of models.

Models	Word error rate (%)
A-WaveNet-CTC	57.6
V-WaveNet-CTC	98
AV-WaveNet-CTC-I	39.4
AV-WaveNet-CTC-A-I-3	**17.8**
AV-WaveNet-CTC-A-I-5	17.9
AV-WaveNet-CTC-A-I-7	17.9
AV-WaveNet-CTC-A-I-10	17.9
AV-WaveNet-CTC-H	25.2
AV-WaveNet-CTC-A-H-3	18.1
AV-WaveNet-CTC-A-H-5	18.0
AV-WaveNet-CTC-A-H-7	17.9
AV-WaveNet-CTC-A-H-10	18.2

For the models with cross-attention in input layer we proposed, the error rates are reduced with relative reduction 4.5% in AV-WaveNet-CTC-A-I-7 ($l = 7$) for Tibetan and 39.8% in AV-WaveNet-CTC-A-I-3($l = 3$) for English compared with A-WaveNet-CTC. Compared with AV-WaveNet-CTC-I, the error rate is reduced by 35.1% in AV-WaveNet-CTC-A-I-7 for Tibetan and by 21.6% in AV-WaveNet-CTC-A-I-3 for English. These show that cross-attention mechanism introduced in input layer of WaveNet-CTC can improve the performance of model for audio-visual speech recognition.

However, we also see that as the range of attention enlarges, the error rate of AV-WaveNet-CTC-A-I model increases, which indicates that the excessive attention range interferes with the alignment of audio and video information.

For the cross-attention applying into the hidden feature fusion, the AV-WaveNet-CTC-A-H models for English also achieve the better performance than AV-WaveNet-CTC-H which directly concatenates two modal hidden features, but they are not better than AV-WaveNet-CTC-A-I. However, for Tibetan, the recognition results of AV-WaveNet-CTC-A-H models are worse than AV-WaveNet-CTC-H. We analyze that the hidden representation transformed from poor visual information disturbs the attention mechanism to find the related visual frames for audio frame. Whatever, from Table 1 and Table 2, we can see that the cross-attention for middle fusion is not better than for early fusion. According to our analysis, the reason maybe is that WaveNet model is composed of dilated causal convolutional layers, where it enlarges the receptive field by skipping input values with a certain step, it is more difficult to correctly align the visual frames with acoustic frame at higher layer.

TABLE 3. CTC output of models for a Tibetan video file.

Models	CTC Output	Single syllable edit distance
Reference	ལུག པགས དངདབྱར རྩྭ དགུན འབུ ཉོ བར འགྲོ གི ཡིན	
A-WaveNet-CTC	ཁྲིམ ལུག སུར དང རྩྭ ཚང སྒོར གཏོང བར འགྲོ གི ཡིན	6
V-WaveNet-CTC	ཡོད	11
AV-WaveNet-CTC-I	ལུག གཙོ མོའི སློབ ཉོ བར འགྲོ ཡིན	7
AV-WaveNet-CTC-A-I-3	ལུག དང རྩྭ དགུན དངུལ ཉོ པར འགྲོ གི	5
AV-WaveNet-CTC-A-I-5	པགས དང སྤངས ཚོ པོ ཉོ བར འགྲོ གི ཡིན	5
AV-WaveNet-CTC-A-I-7	གི ལུག པགས དང རྒྱ དགུན འབུ དངུལ ཉོ བར འགྲོ གི ཡིན	4
AV-WaveNet-CTC-A-I-10	སྡུག པགས དང མཇུག མང གི ཡིན	8
AV-WaveNet-CTC-H	དེང པོ འབུ ཉོ འགྲོ གི ཡིན	6
AV-WaveNet-CTC-A-H-3	NULL	11
AV-WaveNet-CTC-A-H-5	ཞོགས དང དགུན འབུ ཉོ འགྲོ ཡིན	6
AV-WaveNet-CTC-A-H-7	ཡོད	11
AV-WaveNet-CTC-A-H-10	NULL	11

TABLE 4. CTC output of models for an English video file.

Models	CTC Output	Word edit distance
Reference	gregory and tom chose to watch cartoons in the afternoon	
A-WaveNet-CTC	gregory and Tom chose to	6
V-WaveNet-CTC	he is like	10
AV-WaveNet-CTC-I	gregory of Tom chose to watch it cartoons in afternoon	4
AV-WaveNet-CTC-A-I-3	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-I-5	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-I-7	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-I-10	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-H	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-H-3	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-H-5	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-H-7	gregory and Tom chose to watch cartoons in afternoon	2
AV-WaveNet-CTC-A-H-10	gregory and Tom chose to watch cartoons in afternoon	2

Table 3 and Table 4 indicate the predicted label sequences outputted from different models for a Tibetan video file and an English video file, respectively. Obviously, in Table 3 the models with cross-attention in input layer have more accurate recognition results, especially the AV-WaveNet-CTC-A-I-7 has the smallest edit distance of single syllable. Although both AV-WaveNet-CTC-A-I and AV-WaveNet-CTC-A-H models got the same edit distance for an English video, the models with cross-attention are still better than audio-only WaveNet-CTC and the model with the conventional feature concatenation in early fusion.

TABLE 5. 10 Epochs training time of AV-WaveNet-CTC-A-I model with different attention range for Tibetan.

Models	Time (s)	Word error rate (%)
AV-WAVENET-CTC-I	2344	56.9
AV-WaveNet-CTC-A-I-3	4640	50.9
AV-WaveNet-CTC-A-I-5	6228	51.6
AV-WaveNet-CTC-A-I-7	8232	50.4
AV-WaveNet-CTC-A-I-10	15794	50.6

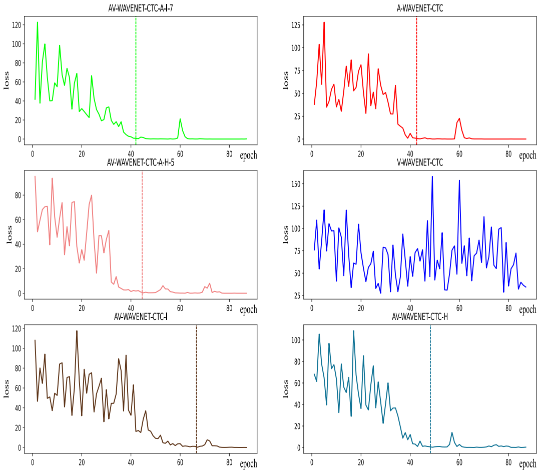


FIGURE 11. CTC loss curves of 6 models for Tibetan. The solid line shows the CTC loss on the training set, the vertical dash line indicates the point of the loss value 0.9. The x-axis represents the training epochs, and the y-axis represents the loss value of the corresponding model.

For the evaluation of computation effectiveness of our method, we compare CTC loss curves of models for Tibetan in training, as shown in Fig. 11, and also demonstrate the 10 epochs training time of AV-WaveNet-CTC-A-H model for English AV-ASR with different attention range on the same computer, as shown in Table 4.

Fig. 11 shows the CTC loss curves for 6 models during training. Green line of AV-WaveNet-CTC-A-I-7 converges the fastest, which has the highest accuracy for Tibetan AV-ASR.

In Table 5, with the half size of sliding window increasing, the 10 epochs training time increases. Especially when attention range changes from 7 to 10, training time has increased tremendously which means the demand for computing resources has also increase a lot. Actually, when attention range is 20, the model cannot run on our computer with Intel Core i7-9700K CPU and two Nvidia RTX 2070 Super GPUs.

VI. CONCLUSION

We present a WaveNet-CTC with cross-attention for feature fusion in AV-ASR. Our method improves performance by assigning different weights calculated by attention mechanism to each frame of input video feature and relating them with acoustic feature frame. Moreover, in the process of calculating the attention coefficient, we only calculate the local attention coefficient which significantly speeds up the process of learning. And by comparing the accuracy of AV-WaveNet-CTC-A-I-*, we show that excessive attention range may cause the accuracy of the model to drop. The choice of the right range of attention is also very important. WaveNet-CTC introduced with cross-attention is more effective for input-layer multimodal feature fusion owing to WaveNet structure.

REFERENCES

[1] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, ''Modality attention for End-to-End audio-visual speech recognition,'' 2018, *arXiv:1811.05250*. Accessed: Aug. 15, 2019. [Online]. Available: http://arxiv.org/abs/1811.05250

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, ''WaveNet: A generative model for raw audio,'' 2016, *arXiv:1609.03499*. Accessed: Oct. 20, 2019. [Online]. Available: http://arxiv.org/abs/1609.03499

[3] J. Luettin, N. A. Thacker, and S. W. Beet, ''Visual speech recognition using active shape models and hidden Markov models,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf. Proc.*, vol. 2. Atlanta, GA, USA, May 1996, pp. 817–820, doi: 10.1109/ICASSP.1996.543246.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, ''Active appearance models,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001, doi: 10.1109/34.927467.

[5] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, ''A comparison of model and transform-based visual features for audio-visual LVCSR,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Tokyo, Japan, Aug. 2001, pp. 825–828, doi: 10.1109/ICME.2001.1237849.

[6] P. S. Aleksic and A. K. Katsaggelos, ''Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5. Montreal, QC, Canada, May 2004, doi: 10.1109/ICASSP.2004.1327261.

[7] H. Sak, A. Senior, K. Rao, and F. Beaufays, ''Fast and accurate recurrent neural network acoustic models for speech recognition,'' 2015, *arXiv:1507.06947*. Accessed: Sep. 06, 2019. [Online]. Available: http://arxiv.org/abs/1507.06947

[8] H. Soltau, H. Liao, and H. Sak, ''Neural speech recognizer: Acoustic-to-Word LSTM model for large vocabulary speech recognition,'' 2016, *arXiv:1610.09975*. Accessed: Sep. 06, 2019. [Online]. Available: http://arxiv.org/abs/1610.09975

[9] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, ''The microsoft 2017 conversational speech recognition system,'' 2017, *arXiv:1708.06073*. Accessed: Sep. 06, 2019. [Online]. Available: http://arxiv.org/abs/1708.06073

[10] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ''Lip reading sentences in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.

[11] S. Petridis, Y. Wang, Z. Li, and M. Pantic, ''End-to-End audiovisual fusion with LSTMs,'' in *Proc. 14th Int. Conf. Auditory-Visual Speech Process.*, Aug. 2017, pp. 1–5.

[12] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, ''Multi-channel attention for End-to-End speech recognition,'' in *Proc. Interspeech*, Sep. 2018, pp. 17–21, doi: 10.21437/Interspeech.2018-1301.

[13] S. Palaskar, R. Sanabria, and F. Metze, ''End-to-End multimodal speech recognition,'' 2018, *arXiv:1804.09713*. Accessed: Dec. 21, 2018. [Online]. Available: http://arxiv.org/abs/1804.09713

[14] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, ''End-to-end audiovisual speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6548–6552, Accessed: Aug. 12, 2020. [Online]. Available: http://arxiv.org/abs/1802.06424

[15] G. Sterpu, C. Saam, and N. Harte, ''Attention-based audio-visual fusion for robust automatic speech recognition,'' in *Proc. Int. Conf. Multimodal Interact. (ICMI)*, Boulder, CO, USA, 2018, pp. 111–115, doi: 10.1145/3242969.3243014.

[16] F. Tao and C. Busso, ''End-to-End audiovisual speech recognition system with multitask learning,'' *IEEE Trans. Multimedia*, early access, Feb. 28, 2020, doi: 10.1109/TMM.2020.2975922.

[17] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, ''Multilingual speech recognition with a single End-to-End model,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4904–4908.

[18] S. Petridis and M. Pantic, ''Deep complementary bottleneck features for visual speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2304–2308.

[19] F. Tao and C. Busso, ''Gating neural network for large vocabulary audiovisual speech recognition,'' *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1290–1302, Jul. 2018, doi: 10.1109/TASLP.2018.2815268.

[20] A. H. Abdelaziz, ''Turbo decoders for audio-visual continuous speech recognition,'' in *Proc. Interspeech*, Aug. 2017, pp. 3667–3671, doi: 10.21437/Interspeech.2017-799.

[21] Y. Zhao, J. Yue, W. Song, X. Xu, X. Li, L. Wu, and Q. Ji, ''Tibetan multi-dialect speech and dialect identity recognition,'' *Comput., Mater. Continua*, vol. 60, no. 3, pp. 1223–1235, 2019, doi: 10.32604/cmc.2019.05636.

[22] N. Harte and E. Gillen, ''TCD-TIMIT: An audio-visual corpus of continuous speech,'' *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015, doi: 10.1109/TMM.2015.2407694.

[23] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, ''Pixel recurrent neural networks,'' 2016, *arXiv:1601.06759*. Accessed: Oct. 20, 2019. [Online]. Available: http://arxiv.org/abs/1601.06759

[24] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks* (Studies in Computational Intelligence), vol. 385. Berlin, Germany: Springer-Verlag 2012.

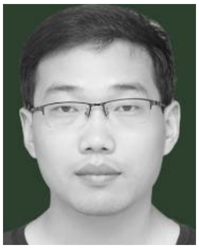
[25] A. Graves and N. Jaitly, ''Towards end-to-end speech recognition with recurrent neural networks,'' in *Proc. 31st Int. Conf. Mach. Learn. (PMLR)*, vol. 32, no. 2, 2014, pp. 1764–1772.

[26] S. Xu. *Speech-to-Text-Wavenet: End-to-End Sentence Level Chinese Speech Recognition Using Deepmind's Wavenet*. Accessed: May 1, 2020. [Online]. Available: https://github.com/CynthiaSuwi/Wavenet-demo

[27] Kim and Park. Speech-to-Text-WaveNet. (2016). *GitHub Repository*. [Online]. Available: https://github.com/buriburisuri/

[28] D. Bahdanau, K. H. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: https://arxiv.org/abs/1409.0473

HUI WANG received the M.S. degree in computer science and technology from Jilin University, in 2002. He currently works as a Full Professor with the Minzu University of China. His research interests include machine learning, data mining, and signal processing.



FEI GAO was born in Hebei, China, in 1992. He received the B.Eng. degree in electronic information engineering from the Hebei University of Science and Technology. He is currently pursuing the M.Sc. degree in computer science and technology with the Minzu University of China. He worked with China Electronics Technology Group Corporation. His current research interests include speech recognition, machine learning, and digital image processing.

YUE ZHAO (Member, IEEE) is currently a Professor in automation engineering with the Minzu University of China. She is the author of three books, more than 50 articles, and more than four inventions. Her research interests include probabilistic graphical models and speech signal processes and applications, computer vision, and embedded systems.

LICHENG WU received the Ph.D. degree in robotics from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1995. He is currently a Full Professor and the Dean of the School of Information Engineering, Minzu University of China. His research interests include speech recognition, artificial robotics, and computer games.