ALIGNING AUDIOVISUAL FEATURES FOR AUDIOVISUAL SPEECH RECOGNITION

Fei Tao and Carlos Busso

Multimodal Signal Processing (MSP) Lab, Department of Electrical and Computer Engineering
University of Texas at Dallas, Richardson TX 75080, USA

fxt120230@utdallas.edu, busso@utdallas.edu

ABSTRACT

Visual information can improve the performance of *automatic speech recognition* (ASR), especially in the presence of background noise or different speech modes. A key problem is how to fuse the acoustic and visual features leveraging their complementary information and overcoming the alignment differences between modalities. Current *audiovisual ASR* (AV-ASR) systems rely on linear interpolation or extrapolation as a pre-processing technique to align audio and visual features, assuming that the feature sequences are aligned frame-by-frame. These pre-processing methods oversimplify the phase difference between lip motion and speech, lacking flexibility and impairing the performance of the system. This paper addresses the fusion of audiovisual features with an *alignment neural network* (AliNN), relying on *recurrent neural network* (RNN) with attention model. The proposed front-end model can automatically learn the alignment from the data. The resulting aligned features are concatenated and fed to conventional back-end ASR systems. The proposed front-end system is evaluated with matched and mismatch channel conditions, under clean and noisy recordings. The results show that our proposed approach can relatively outperform the baseline by 24.9% with *Gaussian mixture model with hidden Markov model* (GMM-HMM) back-end and 2.4% with *deep neural network with hidden Markov model* (DNN-HMM) back-end.

Index Terms— audiovisual speech recognition, deep learning, recurrent neural network, attention model

1. INTRODUCTION

Current *automatic speech recognition* (ASR) systems have reached level of performance that are high enough for practical human-computer interfaces, as demonstrated by commercial products such as Siri, Google voice assistant, Alexa and Cortana. As the acoustic noise in the environment increases, these systems become less effective. It is important to explore solutions to improve the robustness of ASR systems that work regardless of the noise condition. Exploring audiovisual solutions is an appealing approach. Previous studies have shown that *audiovisual automatic speech recognition* (AV-ASR) can overcome the drop in performance caused by noise [1, 2].

A key open challenge is to find effective approaches to combine audio and visual features, as inappropriate fusion schemes may impair the system performance [3, 4]. Conventional approaches include fusing the likelihood scores of modality-dependent systems (i.e., decision level fusion), concatenating the audiovisual features (i.e., feature level fusion), or creating audiovisual latent variables within the models to capture the relationship between modalities (i.e., model-level fusion). In most cases, the audiovisual features are assumed to be synchronized. Since their sampling rate are often different, the features are interpolated or extrapolated to align the modalities. However, audiovisual features are not synchronized. Previous work suggests the existence of a time-variant phase between the audiovisual streams. Studies have reported differences of hundred of milliseconds between modalities [5, 6, 7], which correspond to more than three video frames. Bregler and Konig [8] estimated the mutual information between speech and lip features by adding temporal offsets, assuming that lip movements precede speech. Their analysis showed that the best alignment was with a shift of 120 milliseconds. The challenge is that this phase is not constant or even consistent. For certain phones the modalities are synchronized, for other speech precedes facial movements, and for other facial movements precede speech [5, 9, 10]. These studies suggest that the modalities have to be aligned within a temporal window.

This study introduces the *alignment neural network* (AliNN), a data-driven front-end framework to learn the time-variant phase between audiovisual modalities. The framework relies on sequence-to-sequence learning with attention model [11, 12]. AliNN creates synchronized audiovisual sequences than can be used as features of a back-end ASR system, learning the alignment rules from the data. This appealing framework addresses one of the key problems in AV-ASR that is often ignored.

The proposed approach is evaluated on the CRSS-4ENGLISH-14 corpus [13]. The results show that using AliNN as a front-end improves the performance of different back-end systems, compared to simple linear interpolation of the audiovisual features. It shows that under the ideal condition, our proposed approach obtains a relative improvement of 24.9% with a back-end implemented with *Gaussian mixture model with hidden Markov model* (GMM-HMM), and 2.4% with a back-end implemented with *deep neural network with hidden Markov model* (DNN-HMM). Under channel mismatched condition or noise condition, the proposed framework obtains even larger performance gains, reaching up to 70.5% relative improvements. The experimental evaluation demonstrates the important of synchronizing the audiovisual modalities, a step that is ignored by current systems. The proposed front-end system is flexible, offering a principle framework to improve AV-ASR systems.

2. RELATED WORK

Adding visual information is an appealing solution to build ASR systems robust against background noise or different speech modes [14, 15, 16]. Neti et al.[14] showed that by combining audiovisual information, the ASR system would have better performance. In addition, Tao et al.[17] and Tao and Busso [13] showed that audiovisual

This work was funded by NSF CAREER award IIS-1453781.

978-1-5386-1737-3/18/$31.00

feature is also helpful for other speech processing systems such as *voice activity detection* (VAD)

The fusion approaches used in previous work are mainly grouped at feature, decision and model level integration [1]. At the feature level, the audio and visual features are concatenated, which tends to lead to lower performance [14]. One common simplification is to align the modalities frame-by-frame, which lead to potential problems [3] as the speech and lip movements are not necessary synchronized [9]. For certain phonemes, there are anticipatory movements that precede speech production [5]. For other phoneme, speech precede lip motions. Decision level fusion combines the likelihood scores from modality-depended systems. These systems also operate with common frame rates so the synchronization problem still play an important role. Model level fusion finds hidden latent variables to better capture the relationship between the modalities [18, 19]. While certain systems can model the asynchrony between modalities, such as *asynchronous hidden Markov models* (AHMMs) [20], the level of asynchrony is limited to reduce the complexity of the system.

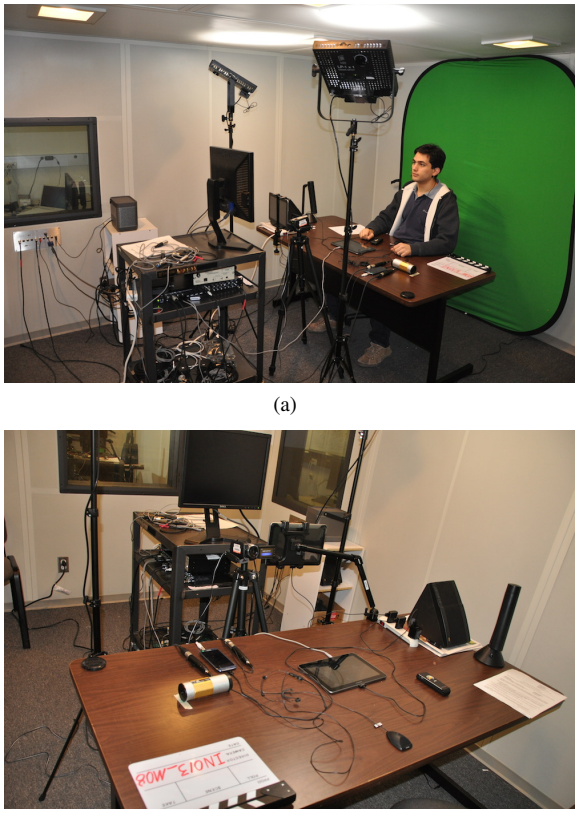
Deep learning techniques have recently emerged as powerful techniques in audiovisual speech processing systems. The front-end for feature extraction and back-end for recognition can be learned from the data, without pre-assumed rules. Ninomiya et al.[21] learned new audiovisual features by extracting a hidden representation with deep bottleneck features. This feature representation was used as audiovisual features of a back-end AV-ASR system. Ngiam et al.[4] proposed multimodal *deep neural network* (DNNs) which can learn the relationship between the modalities from the data. Tao et al.[13] extended the work creating bimodal *recurrent neural network* (RNN), capturing the temporal information with recurrent connections. Chung et al.[22] proposed a bimodal end-to-end framework relying on RNN with attention model to capture temporal dependencies. Sanabria et al.[23] used *connectionist temporal classification* (CTC) as loss function in building an end-to-end system to learn text sequence from audiovisual representation sequences.

Previous studies have often relied on feature interpolation or extrapolation to form audio and visual feature vectors with the same length [1, 24, 14]. This pre-processing step is essential to align audiovisual feature sequences. However, this approach does not provide the flexibility to capture the time-variant phase between modalities for different phonetic units. Contextual features used in ASR systems implemented with DNN [25] can help to model temporal information by concatenating multiples previous and future frames [26]. However, the systems does not provide enough flexibility by defining a fixed window. Bahdanau et al.[11] propose to align the input and output sequences with different lengths using RNN with attention model in the context of machine translation. Our study explores this solution for audiovisual ASR, aligning audiovisual features within a data-driven front-end framework based on attention model. As demonstrated in this study, this front-end approach leads to measurably improvements in *word error rate* (WER).

3. DATABASE AND AUDIOVISUAL FEATURES

3.1. CRSS-4ENGLISH-14 corpus

This study uses the CRSS-4ENGLISH-14 corpus [13]. The corpus was collected in a $13ft \times 13ft$ ASHA certified sound booth. The sound booth was illuminated by two LED light panel during the data collection, providing uniform illumination (Fig. 1(a)). The corpus includes recordings from 442 subjects (217 females and 225 males) from four English accents: Australian (103), Indian (112), Hispanic (112) and American (115). The data was manually transcribed.



(a)

(b)

Fig. 1. Data collection setting with the equipments used in the recordings of the CRSS-4ENGLISH-14 corpus.

The data contains audio and video information using multiple recordings (five microphones and two cameras). For the audio, this study uses a close-talking microphone (Shure Beta 53) and a microphone from the tablet placed about 150 centimeters from the subject (Samsung Galaxy Tab 10.1N). The two microphones collect audio at 44.1 kHz. For the video, this study uses a *high definition* (HD) camera (Sony HDR-XR100) and the tablet's camera. The videos of HD camera was set to 1440×1080 resolution and a sampling rate of 29.97 *frames per second* (fps). The videos from the tablet was set to 1280×720 resolution and 24 fps. A monitor was placed in front of the speaker to show the required tasks. The equipments are shown in Figure 1(b). A clapping board was used at the beginning of each recordings to synchronize all the channels.

The tasks for the subjects included two sections: clean and noisy recordings. In the clean recordings, the protocol included read and spontaneous speech. The read speech task required the speaker to read prompted text shown in the monitor, including isolated words, short phrases, and complete sentences. The spontaneous speech task required the speaker to answer questions shown on the monitor. In the noisy recordings, a portion of the read text was repeated by randomly selecting slides used during the clean recordings (i.e., the material used in the noisy section was a subset of the material used in the clean section). The difference is that background noise was played through a loud speaker (Beolit 12) in the sound booth to emulate noisy environments.

This study only uses the set with American accent to reduce

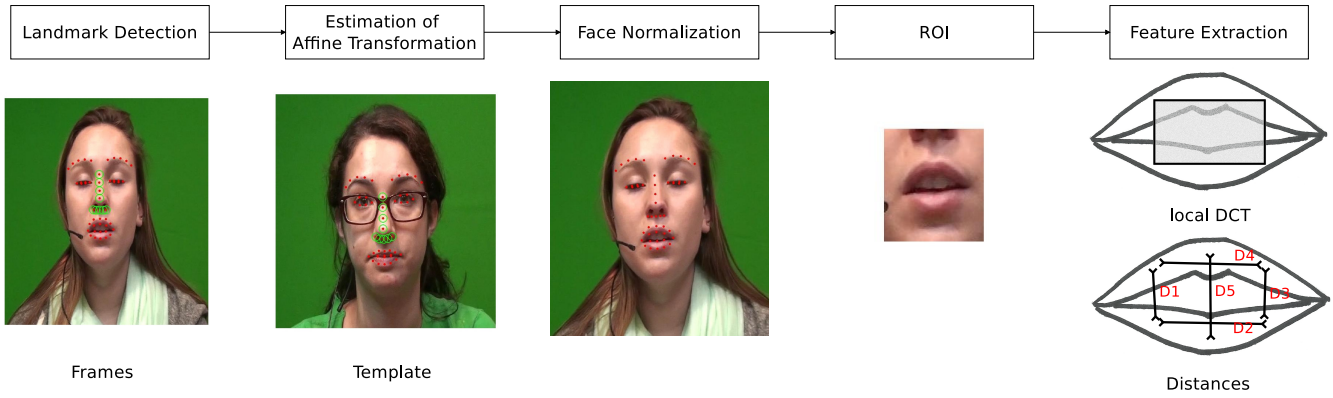


Fig. 2. The procedure to extract visual features. A frontal face image was selected as a global template. Points with green circles were defined as rigid points, used to normalize the size and pose of the face. After detecting the ROI, we estimate 5D geometric distances between landmarks and 25D local DCT, extracted from the gray box (30D visual feature vector).

intra-speaker variability due to differences in accent. The duration of the set is 60 hours and 48 minutes, including 55 females and 50 males. We separate the data into train (70 speakers), develop (10 speakers) and test (25 speakers) sets. We did not use data from 10 speakers, because some of the videos were not properly stored.

3.2. Audiovisual Features

This study uses the audiovisual features proposed by Tao and Busso [16]. The audio is downsampled to 16 kHz. We extract 13D *Mel frequency cepstrum coefficients* (MFCCs), which are used as acoustic features. We use a 25 milliseconds window shifted by 10 milliseconds, creating 100 fps.

The visual features correspond to geometrical and appearance features extracted from the *region of interest* (ROI) centered around the lips. Figure 2 shows the details of the visual extraction process. We use IntraFace [27] to automatically extract 49 facial landmark. A frontal face image from one subject is selected as the global template. This template is used to normalize the size and pose of the face for each frame. We estimate this affine transformation by comparing selected rigid points in both the input frame and the template. These rigid points are selected since they are less sensitive to facial movements (green dots in Figure 2). We define the ROI, after normalizing the face. We estimate a 5D geometric vector describing distances between landmarks in the lips (see Fig. 2). We define a box inside the mouth from which we extract 25D *discrete cosine transform* (DCT) coefficients. The final vector is a 30D feature vector after concatenating the geometric and DCT features.

4. PROPOSE APPROACH

4.1. Alignment Neural Network

We propose the front-end *alignment neural network* (AliNN), which aims to learn the synchrony between audiovisual modalities from the data. The key idea is to synchronize the visual features with the acoustic features, providing a new aligned visual sequence with the same length as the acoustic features. The alignment between modalities is learned from the data. Figure 3 shows the proposed deep learning structure, which consists of a dynamic part and a static part.

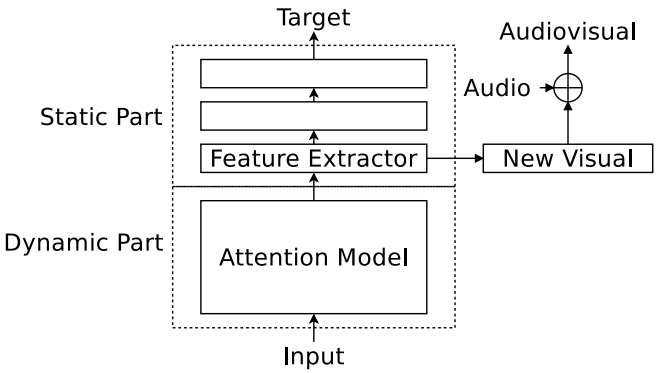


Fig. 3. Diagram describing the proposed AliNN front-end system. It includes a dynamic part, implemented with RNN and attention model, and a static part, implemented with DNN with static layers. The first layer of the static part extracts audiovisual features.

The dynamic part relies on attention model to apply sequence-to-sequence learning. The attention model was introduced by Bahdanau et al.[11] and Chorowski et al.[12]. Figure 4 shows the core part of the attention model. The output state is modeled based on the previous state and the linear combination of all hidden values in the input layer. Very importantly, the output and input sequences can have different lengths. Equations 1 and 2 show the linear combination equations, where i is the timestep in the output layer, and j is the timestep in the input layer. h_j is the hidden value at the timestep j in the input layer, a_{ij} is the weight of the input frames at the timestep i of the output, and c_i is the linear combination of all hidden values in the input layer for the output timestep i. Equation 2 shows that the summation of all the weights for the output at timestep i is equal to 1. Each output frame is a linear combination of the input frames. The weight a_{ij} plays the role of aligning the input at timestep j and the output at timestep i. Therefore, the attention model can learn the alignment between sequences even though they have different lengths.

$$\mathbf{c}_i = \sum_{j=1}^{T} a_{ij} h_j \tag{1}$$

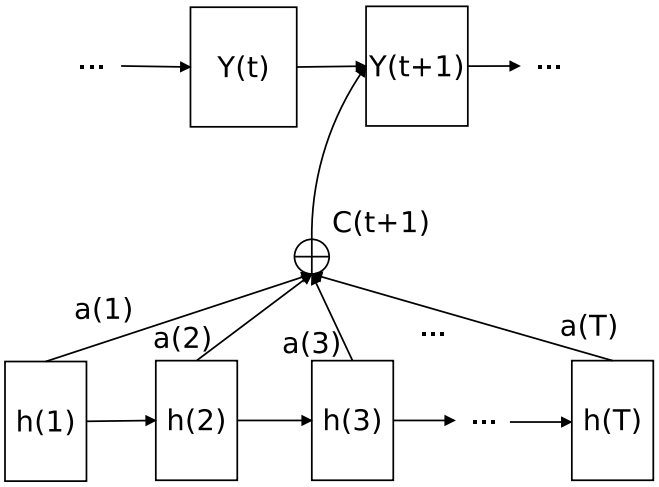


Fig. 4. Diagram of the attention model. The input and output sequences are unrolled along the time axis. The state in the output layer is controlled by the previous state and the linear combination of all the hidden values in the input layer. The input and output sequences can have different lengths.

$$\sum_{j=1}^{T} a_{ij} = 1 \tag{2}$$

The attention model offers a principled approach to synchronize audiovisual features. The formulation in AliNN takes visual features as inputs and acoustic features as the target outputs. The training is equivalent to regress visual features to acoustic features, where the *mean squared error* (MSE) is used as the loss function. The layers on the dynamic network of the AliNN framework are implemented with *long short-term memory* (LSTM) networks.

The AliNN front-end does not consider the phonetic content in the speech. The static network on top of the dynamic layers aims to predict the acoustic features after the alignment. The static part is implemented with several layers without recurrent connections. The first layer of the static part serves as a feature extractor, providing a new visual feature representation that is timely aligned with the acoustic features (see Fig. 3). Once the AliNN framework is trained, the output of the first layer of the static part is used as our new visual feature vector, which is concatenated with the original acoustic feature to form the final audiovisual feature vector. This feature vector is then used as the input of our back-end ASR recognizer.

The structure of the dynamic part is implemented with two recurrent layers (i.e., LSTM) with 30 neurons in each layer. The static part has three fully connected layers with *rectified linear unit* (ReLU) as activation function. Each layer has 30 neurons. The number of neuron in each layer is the same as the dimension of the original visual feature vector, so the dimension of the new visual feature is not changed.

4.2. Training AliNN

The asynchrony between audiovisual features reported in previous studies is around 100 milliseconds [7, 8]. Therefore, we restrict the search during the training process by segmenting the sequences with one second window shifted by 0.5 seconds (i.e., 0.5 seconds overlap between two contiguous sub-sequences). Zero-padding is added if the sub-sequences are less than one second. Segmenting the sequences also makes the training process more efficient, reducing the time to train the AliNN system.

The AliNN front-end is trained with the sub-sequences of audiovisual features. Hidden representations are also extracted with the sub-sequences. We combine the outputs obtained from the feature representation layers by averaging the overlap from contiguous subsequences. With this approach, we obtain a new feature sequence with the same length as the unsegmented target sequence. This process facilitates the training of the AliNN front-end, providing a new synchronized visual vector that has the same length, and is timely aligned with the acoustic features. The back-end ASR system can be trained by concatenating the new aligned visual features with the acoustic features.

5. EXPERIMENTAL EVALUATION

5.1. Experiment Settings

We evaluate the proposed approach on the CRSS-4ENGLISH-14 corpus. The AliNN approach is used as a front-end of two back-end ASR systems implemented with GMM-HMM and DNN-HMM. For the GMM-HMM back-end, the feature vector is augmented with delta and delta-delta information. For the DNN-HMM, we used the contextual features with a window of 15 frames (i.e., seven previous frames and seven future frames). The *word error rate* (WER) is used as the performance metric.

For training the models, we use data collected from the close-talking microphone and HD camera using clean recordings. For testing the models, we have two channel conditions. The *ideal* channel condition uses audio from the close-talking microphone and videos from the HD camera. Since the back-end systems are trained with data collected with these devices, we have matched channel conditions when we test the models with the ideal condition. The second condition is referred to as the *tablet* channel condition, where the audio and video are recorded with the tablet. Testing the model with the tablet condition creates channel mismatched, so we expect lower performance. Since the videos from the HD camera and tablet camera have different sampling rates, we linearly interpolated the 30D features obtained from the tablet videos, matching the sampling rate of the HD camera. This step is implemented before they are fed into AliNN. We evaluate the models with clean and noisy recordings (Sec. 3.1).

The front-end for the baseline system consists of linear interpolation of the original visual features to match the sampling rate of the acoustic features, which is a common approach used in AV-ASR. After interpolation, we concatenate the audiovisual features. While the baseline systems used the interpolated version of the visual features, the proposed method uses the aligned visual features which have the same sampling rate of the audio stream, and are timely aligned with the acoustic features. The acoustic features are the same for the baseline and the proposed front-end system.

5.2. Results and Discussion

Table 1 presents the results of the experimental evaluation. It shows that the AliNN front-end can outperform linear interpolation method with either GMM-HMM or DNN-HMM back-end system under most conditions, showing the capability of our proposed front-end framework.

With a GMM-HMM back-end under ideal channel conditions, the proposed front-end provides relative improvements over the

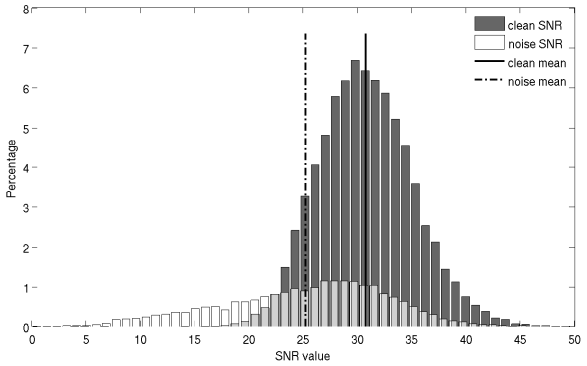
Table 1. WER of systems implemented with linear interpolation (baseline denoted by *LInterp*) and the proposed front-end framework (denoted *AliNN*). The table shows results with matched and mismatched channel conditions, with clean and noisy conditions, and with two back-end ASR systems.

Front-end	MODEL	Ideal Conditions		Tablet Conditions	
		Clean WER	Noise WER	Clean WER	Noise WER
LInterp	GMM-HMM	23.3	24.2	24.7	**30.7**
AliNN	GMM-HMM	**17.5**	**19.2**	**22.7**	35.6
LInterp	DNN-HMM	4.2	4.9	15.5	15.9
AliNN	DNN-HMM	**4.1**	**4.5**	**4.6**	**10.0**

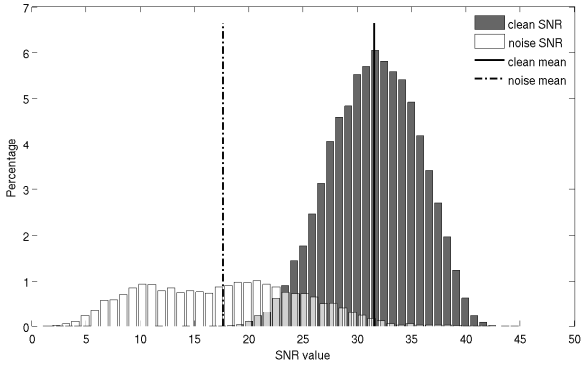
baseline of 24.9% under clean recording, and of 20.7% under noisy recordings. Under tablet channel condition, the proposed approach is 8.1% (relative) better than the baseline in clean recordings. However, the approach is 16.0% (relative) worse than the baseline in noisy recordings. The improvement under ideal channel conditions is larger than the tablet channel condition. A reason for this result is the use of linear interpolation to upsample the visual features from 24 fps to 29.97 fps before feeding the visual features into the AliNN front-end. This linear interpolation step may impair the alignment with acoustic features.

With a DNN-HMM back-end, the proposed approach outperforms the baseline under all conditions. Under ideal channel conditions, the AliNN front-end obtains a relative improvements of 2.4% for clean recordings, and 8.2% for noisy recordings. Notice that the DNN-HMM framework provides better performance than the GMM-HMM framework with WER below 5% for the ideal channel condition. Therefore, it is harder to improve the results. Under tablet channel condition, the proposed approach achieves impressive relative gains over the baseline front-end (70.5% for the clean recordings and 39.7% for noisy recordings).

We also observe that the performance in noisy recordings under tablet channel condition is lower than the ideal tablet conditions. This result is explained by the setup used to collect the corpus. The load speaker playing the noise was physically closer to the tablet microphone than the close-talking microphone. As a result, the level of noise was higher in the microphone from the tablet. Figure 5 presents the distribution for the *signal-noise-ratio* (SNR) for the ideal and tablet channel conditions, automatically estimated using the NIST speech SNR tool [28]. The figure shows that under the ideal channel condition, the clean and noisy recordings have important overlap, so the WER by the system is less affected by the noise. Under the tablet channel condition, however, the noisy recordings have significant lower SNR. Therefore, the ASR task is more challenging.



(a) SNR for close-talk microphone



(b) SNR for tablet microphone

Fig. 5. The SNR distribution for close-talking and tablet microphones. The gray bars represent the noisy recordings, and the white bars represent the clean recordings.

6. CONCLUSIONS AND FUTURE WORK

This paper proposed the *alignment neural network* (AliNN), which is a principled front-end framework to model the temporal alignment between audiovisual features. AliNN learns the alignment from the data without the need of labels (e.g., transcription). The framework uses attention models with LSTM, where the task is to estimate acoustic features from visual features. The attention model framework addresses the difference in sampling rate in the modalities, creating a new visual vector that not only has the same length as the acoustic feature vector, but also is timely synchronized with speech.

The experimental evaluation conducted on the CRSS-4ENGLISH-14 corpus shows that the proposed front-end provides better performance than the baseline front-end (matching the sampling rate using linear interpolation on the visual features). We evaluated the front-end approaches with back-end ASR systems implemented with either GMM-HMM or DNN-HMM. The results show that the AliNN approach can outperform the baseline front-end under most of the conditions (channel mismatched condition, noise versus clean conditions). The results are specially impressive with DNN-HMM, where the WER are consistently reduced across all conditions when using the proposed AliNN front-end. Our approach can model the temporal and spatial relationship between audio and visual modalities directly from data without pre-defined rules, which addresses one the challenges in AV-ASR that is commonly ignored.

There are many directions to extend this work. One of the most unique features of the proposed approach is that it can be used with any back-end system. In the future, we will explore an end-to-end back-end system, which have emerged as a powerful deep learning solution for AV-ASR. We are planning to integrate the AliNN with an end-to-end back-end ASR, providing a unified solution that aligns the audiovisual features, maximizing their phonetic discrimination.

7. REFERENCES

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audiovisual automatic speech recognition," in *Audio-Visual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds., pp. 193–247. Cambridge University Press, Cambridge, UK, February 2015.

[2] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, September 2015.

[3] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. to appear, 2018.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International conference on machine learning (ICML2011)*, Bellevue, WA, USA, June-July 2011, pp. 689–696.

[5] T.J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[6] C. Benoît, "The intrinsic bimodality of speech communication and the synthesis of talking faces," in *The Structure of Multimodal Dialogue II*, M.M. Taylor, F. Néel, and D. Bouwhuis, Eds., pp. 485–502. John Benjamins Publishing Company, March 2000.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[8] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994)*, Adelaide, Australia, April 1994, vol. 2, pp. 669–672.

[9] V. Van Wassenhove, K. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, 2007.

[10] J.X. Maier, M. Di Luca, and U. Noppeney, "Audiovisual asynchrony detection in human speech," *Journal of Experimental Psychology Human Perception and Performance*, vol. 37, no. 1, pp. 245–256, February 2011.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Workshop track at International Conference on Learning Representations (ICLR 2015)*, San Juan, Puerto Rico, May 2015, pp. 1–15.

[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 577–585.

[13] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1938–1942.

[14] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Technical Report 764, Workshop 2000 Final Report, October 2000.

[15] X. Fan, C. Busso, and J.H.L. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *European Signal Processing Conference (EUSIPCO-2011)*, Barcelona, Spain, August-September 2011, pp. 1500–1503.

[16] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Interspeech 2014*, Singapore, September 2014, pp. 1154–1158.

[17] F. Tao, J.H. L. Hansen, and C. Busso, "Improving boundary estimation in audiovisual speech activity detection using Bayesian information criterion," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2130–2134.

[18] S. Gergen, S. Zeiler, A. Abdelaziz, R. Nickel, and D. Kolossa, "Dynamic stream weighting for turbo-decodingbased audiovisual ASR," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2135–2139.

[19] A.H. Abdelaziz, "Turbo decoders for audio-visual continuous speech recognition," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 3667–3671.

[20] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems (NIPS 2002)*, Vancouver, BC, Canada, December 2002, pp. 1237–1244.

[21] H. Ninomiya, N. Kitaoka, S. Tamura, and Y. Iribeand K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 563–567.

[22] J.S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian Conference on Computer Vision (ACCV 2016 Workshop)*, C.S. Chen, J. Lu, and K.K. Ma, Eds., vol. 10117 of *Lecture Notes in Computer Science*, pp. 251–263. Springer Berlin Heidelberg, Taipei, Taiwan, November 2016.

[23] R. Sanabria, F. Metze, and F. D. L. Torre, "Robust end-to-end deep audiovisual speech recognition," *CoRR*, vol. abs/1611.06986, November 2016.

[24] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[25] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, January 2012.

[26] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, July 2017.

[27] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "IntraFace," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[28] V. M. Stanford, "NIST speech SNR tool," https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements, December 2005.