MULTIMODAL SPEAKER ADAPTATION OF ACOUSTIC MODEL AND LANGUAGE MODEL FOR ASR USING SPEAKER FACE EMBEDDING

Yasufumi Moriya *Gareth J. F. Jones*

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland

ABSTRACT

We present an investigation into the adaptation of the acoustic model and the language model for automatic speech recognition (ASR) using speaker face for transcription of a multimedia dataset. We begin by overviewing relevant previous work on the integration of visual signals into ASR systems. Our experimental investigation shows a small improvement in word error rate (WER) for the transcription of a collection of instruction videos using adaptation of the acoustic model and the language model with fixed-length face embedding vectors. We also present potential approaches to integrating human facial information, and body gestures into ASR as further directions for research on this topic.

***Index Terms*—** multimodal speech recognition, face embedding, adaptation

1. INTRODUCTION

Grounding visual information into language understanding is a natural human ability. It has been demonstrated that speech perception and sentence comprehension can be affected by visual context [1, 2], and that the presence of human faces facilitates speech comprehension [3]. Conventional ASR systems only rely on an audio signal, even if a visual signal is available in the data being recognised. Of particular interest in our work is ASR for multimedia data such as user-generated content which is often challenging, since such data is less controlled than standard ASR datasets. High word error rates (WERs) of more than 30-40% have been reported in existing investigations of ASR for multimedia data [4]. On the other hand, several previous investigations have shown that integration of multimodal information into the ASR process can improve WER for multimedia data [5, 6], which encourages us to pursue further work in this direction.

In this paper we investigate the use of speaker faces present in a visual signal for adaptation of an acoustic model and a language model for ASR. Our work is motivated by the multimodal nature of human language processing, and by the recent success of the use of visual information for ASR. In this work, human faces are represented as fixed-length vectors (referred to as "speaker face embedding") extracted using a convolutional neural network (CNN). We hypothesise that ASR can implicitly learn speaker demographic information, speaker emotion, and personal use of language (e.g., choice of words) from speaker face embedding. The adapted acoustic and language models are applied to a multimedia dataset. The key contribution of this work is adaptation of the acoustic model and the language model for ASR using speaker face embedding, and discussion of potential methods to integrate facial expressions and body movement into the ASR process.

The reminder of the paper is organised as follows. Section 2 reviews relevant existing work and identifies the contributions of this work, Section 3 presents our method to integrate facial features into ASR. Section 4 describes the dataset and system configurations for our experiments. Section 5 summarises experimental results, while Section 6 concludes.

2. RELATION TO PRIOR WORK

Multimodal ASR in the form of audio-visual speech recognition (AVSR) is a long-standing research topic. Much of this work has focused on the fusion of lip movement with audio features to enable more robust ASR [7, 8]. However, this approach is limited since construction of large-scale AVSR datasets is not straightforward due to the requirement of precise alignment of lip movement with phonemes and front view of the speaker's face.

More recent work on multimodal ASR has broadened the scope of this topic to remove the constraint of focusing on tracking the speaker's lip movement. Fleischman and Roy [5] adapted a language model using event patterns of a baseball match for recognition of baseball commentary. They represented a sequence of visual and audio contextual signals (e.g., "pitching", "running", and "excited speech") in a codebook, and adapted a language model using Gibbs sampling. They report a high baseline WER of 80.3%, most likely caused by noisy audio conditions. By contrast, their multimodal approach reduced WER to 76.6%. Gupta et al. [6] investigated the use of object features and scene features for adaptation of a recurrent neural network (RNN) language model and a deep neural network (DNN) acoustic model. For each utterance, a video frame randomly chosen from a time range of the utterance was transformed into a fixed-length vector representing object or scene features through a CNN model.

The adapted models were evaluated on a collection of instruction videos. They found that the "visual context" led to reduction in WER. The method was particularly effective on a video where a speaker was talking outside a building with background noise, indicating that the use of scene features implicitly contributes to denoising of acoustic environments. Their recent work incorporates those visual features into an end-to-end speech recognition system [9]. Finally, it has also been found that contextual information such as video titles can effectively augment an RNN language model for ASR [10]. Similar to this previous work, our model adaptation using speaker face embedding does not assume hard alignment of lip features with phonemes.

Acoustic model adaptation is a well-studied topic. In particular, adaptation of DNN acoustic models often employs a speaker specific vector known as an "i-vector". Extraction of i-vectors involves training a total variability matrix, assuming all of the utterances belong to different speakers [11]. For example, speaker level i-vectors can be extracted, and concatenated with acoustic features [12]. Linear feature shift can be computed by transforming speaker-level i-vectors along with acoustic features using a DNN [13]. Normalisation parameters can be estimated with speaker-level i-vectors, speaker clusters computed for test utterances, and the parameters applied to cluster-level i-vectors [14]. Unlike DNN acoustic model adaptation using i-vectors, our approach does not assume speaker labels are available. An off-the-shell face embedding extractor pre-trained on a large amount of labelled data is sufficient to discriminate between faces, see Section 3.

Fusing visual information about human faces and bodies with auditory features has also been explored in existing work. In the field of multimodal affective computing, body gestures, facial expressions, and head movement have been integrated with audio features into affect analysis [15]. Our work focuses on the use of speaker face embedding. We plan to explore more types of features from humans present in a visual stream in future work. Bredin and Gelly [16] propose the use of talking face detection, and face clustering for speaker diarization. They initialise speech clusters with corresponding face clusters, or detected talking faces. While in their study they demonstrated reduction in diarization error rate compared to the audio only baseline, we investigate the use of face embedding for ASR instead of diarization.

The closest existing work to our investigation was conducted by Miao et al. [13]. They extracted speaker attributes (i.e., "age", "gender", and "race") from a visual signal, and combined these attributes with audio features. The difference between this work and ours is that we employ a face embedding extractor and do not rely on categorical values. We hypothesise that extracted fixed-length face embedding vectors implicitly carry these speaker attributes along with speaker identities. Unlike textual corpora, speaker information can be obtained from a visual or an audio stream of multimedia data. We adapt not only the acoustic model, but also the language model, as the vocabulary size and the choice of words can be dependent on age and gender. In recent work, Vanmassenhove et al. found that incorporating gender tags into neural machine translation improves translation of English into several languages [17].

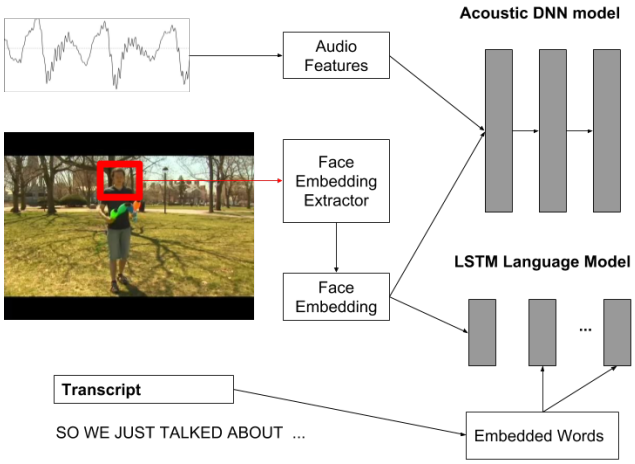

Fig. 1. Adaptation of a DNN acoustic model and an LSTM language model with face embedding. A human face tracked in a video is transformed into a fixed-length vector representation. The weights in grey boxes are updated during training.

3. METHODOLOGY

This section describes extraction of face embedding vectors from videos, our baseline DNN acoustic model and recurrent neural network (RNN) language model with long short-term memory (LSTM), and adaptation of the acoustic model and the language model.

Figure 1 shows our framework for the adaptation of the acoustic model and the language model with speaker face embedding. Extraction of face embedding from each video consists of three steps:

1. a shot boundary detector segments a video based on transitions of different visual patterns.
2. face detection and tracker identify human faces present in detected shots.
3. face embedding is extracted from each face track using a pre-trained CNN model.

As mentioned in Section 2, a CNN model pre-trained on a vast amount of face labels is used to extract face features that are useful to distinguish between other faces. In practice, *pyannote-video*[1] was employed for face embedding extraction. More details on the *pyannote-video* tool can be found

[1]https://github.com/pyannote/pyannote-video

in [16]. The pre-trained CNN model has a residual network architecture trained on roughly 3 million faces with 7,485 labels, and is available as part of a *dlib* model [18, 19].

The baseline DNN acoustic model is a feed-forward neural network with a time-delay architecture [20]. Time-delay neural networks can capture long-term contexts better than simple feed-forward neural networks. During the training phase, the network predicts probabilities of context-dependent phones, given an audio feature vector. Cross-entropy error is computed by comparing the probabilities against the actual phone label. The gradients of the error function are used to update weights of the network hidden layers with the back-propagation algorithm.

When adapting the DNN acoustic model with speaker face embedding, an audio feature vector is concatenated with a speaker face embedding, as in Figure 1. This adaptation method is similar to that in [12], except that the i-vector is replaced for face embedding. While adaptation of the acoustic model with an i-vector requires manually annotated speaker labels to generate speaker-level i-vectors, adaptation with speaker face embedding does not need such labels. Our hypothesis is that face embedding entails speaker attributes, emotion and identities, which can be related to voice pitch, voice tone, accent, pronunciation, and other personal characteristics of speech production.

The baseline neural language model is a RNN language model with LSTM cells [21, 22]. The RNN language model is capable of learning long-term context. However, accumulating gradients of a long sequence may cause the exploding or vanishing gradient problem. The LSTM cells can solve this issue by controlling how much information is retained while reading a sequence. On training the LSTM language model, the model predicts probabilities of the next word given an embedded word and the sequence history. Cross entropy loss is accumulated through an input sequence, and its gradients propagated through the whole sequence to update layer weights with the back-propagation through a time algorithm.

As shown in Figure 1, speaker face embedding is read by the LSTM language model as background context, before the model takes word embedding input. Since face embedding is expected to carry speaker attributes and identities, our hypothesis is that the language model is tuned for gender, age, and speaker specific word choices. The adapted language model can be used to re-rank n-best ASR hypotheses.

4. EXPERIMENTAL SETUP

This section presents the audio-visual dataset used for our experiments and configuration of the models.

4.1. How2 Dataset

The How2 Dataset is a collection of instructional videos [23]. This corpus covers a variety of topics for viewers (e.g., cooking, tennis), and speakers mostly talk in a formal style. Some speakers are present outside a building, and background noise may be present in audio files.

While this corpus was used for the experiments in [6], we applied our own pre-preprocessing, since the original corpus contained noise (e.g., symbols and numbers in transcripts). Therefore, the results of this paper are not directly comparable to the above paper. The cleaning steps aimed at: (1) normalisation of symbols and numbers in transcripts, and (2) segmentation of audio files, and rejection of mismatches between audio files and corresponding transcripts. For more details for the segmentation and cleaning, see [24].

We realised that only one speaker talks in each video. For this reason, we assume that the longest face track detected in each video was a speaker face. Since face detection is computationally demanding, face tracking was initially applied to every 10 seconds of each video to run the algorithm quickly. Then, the face detector was applied again to every frame of the videos in which no faces were detected in the first attempt, since this results in more accurate detection. Face embedding vectors belonging to the longest track of each video were averaged to form a speaker face embedding vector. The vector was substituted for a zero vector, when no face was detected through face detection due to their being no face present or to poor resolution causing a face detector failure. Overall, a zero vector was used for 27 utterances.

For our experiments, the training set of data consists of roughly 107 hours of audio, and the test set 4.8 hours. The decoding graph was trained on the whole training portion of the original dataset (173,684 utterances) using a 3-gram language model with modified Kneser-Ney interpolation, implemented in the SRILM toolkit [25, 26]. When training the LSTM language model, only the training text (34,333 utterances) corresponding to the audio data was used, since face embedding was not prepared for the whole dataset. During the training of the language model, its performance was monitored based on perplexity on the development split of the original corpus (1,852 utterances). The learning rate was divided by 4, when perplexity of the previous epoch was lower than the current epoch. The number of epochs was set to 50. This configuration previously produced the best results in [10]. The model which produced the lowest perplexity during the training was retained as the final model. The number of utterances in the test split of the corpus was 1,680.

4.2. Model Configuration

The audio features used to train the DNN acoustic model was a 40 dimensional filter bank with a window length of 25 ms, and 10 ms frame shift. Speaker face embedding had 128 dimensions, so that concatenation with the audio feature formed a 168 dimensional vector. The time-delay neural network consists of 6 layers with 1024 hidden units each. The output size was 1568 context-dependent phones for the How2 cor-

Table 1. Perplexity (PPL) and WER results for the adaptation of the acoustic model and the language model. "baseline" is the acoustic model without adaptation, "face_emb" is the one with adaptation, "i-vector" uses i-vector adaptation, and "i-vector+face_emb" is adapted with both i-vector and face embedding. "LSTM rerank" is the results for n-best re-ranking produced by the "baseline" system. "LSTM face_emb" is the adapted version of the LSTM language model. "oracle" is the best achievable WER from a decoding graph.

	PPL	WER
baseline	-	22.65
face_emb	-	22.48
i-vector	-	22.66
i-vector+face_emb	-	22.60
LSTM rerank	119.84	21.65
LSTM face_emb	120.44	21.48
oracle	-	17.23

pus. The mini-batch size was 512, and the number of training epochs was 8. The model was built using Kaldi [27].

For comparison to speaker face embedding, we applied i-vector adaptation to the acoustic model. A Guassian mixture model of 1024 components and a total variability matrix were trained on regions of audio where voice activity detection identified speech. i-vectors of 128 dimensions were extracted from 1.5 sec. windows with 0.75 sec. shift. Assuming each video has one speaker, speaker i-vectors were computed by taking average of all the utterances of each video.

The LSTM language model consisted of two layers with 512 hidden units. Word embedding size was 128, which is identical to that of speaker face embedding. The output vocabulary size was 21,336 words. The initial learning rate was 20. The model was trained with 50 epochs. Dropout rate 0.2 was applied to the LSTM layers and word embedding model to avoid over-fitting. The mini-batch size was set to 100. 30-best hypotheses were generated by the baseline ASR for the dataset. The LSTM language model was built with Pytorch [2].

5. RESULTS

Table 1 summarises our experimental results using the adaptation technique proposed in this paper with comparison to non-adapted models. Adaptation of the both acoustic model and language model brought a small gain in WER. As mentioned in Section 4.1, the How2 dataset presents only one speaker per video. Thanks to this consistency, application of the proposed technique was straightforward, and both the models could learn person identities from the training set. Surprisingly, i-vector adaptation and combining i-vector with face embedding did not reduce WER. This is possibly because the acoustic environment of the user generated content is not always clean, and an approach to refining i-vectors such as [13] maybe required for improvement.

[2]https://pytorch.org/

6. CONCLUSION AND FUTURE WORK

This paper investigates the use of face embedding for adaptation of the DNN acoustic model and the LSTM language model. Grounding multimodal information into ASR is motivated by studies on human behaviour and recent works on multimodal ASR, in which integration of the visual (non-lip movement) information into ASR has been established as a promising emerging topic of research. To develop this area of work more attention needs to be devoted to understanding how humans exploit visual cues for speech and language comprehension, as well as how automatic systems can be improved with these cues.

Future work is to refine speaker face embedding. Firstly, face embedding was only extracted from the part of the How2 corpus. This can be extended to the full dataset (173,684 utterances) to create better mapping between speaker faces and speech characteristics. Secondly, using all of the face tracks detected in a video may also lead to better features, rather than merely averaging face embedding vectors of the longest track for each video. Thirdly, temporal features such as face movement and body gestures could be integrated into the ASR systems. Intuitively, there should be high correlation between body movement and speech tone (e.g., opening a mouth wide, while shouting). Such information could be useful for recognition of emotional speech. Finally, we expect speaker face embedding to be more effective on a speech corpus containing a diverse speaker set including children, and on a corpus of a non-English language, where grammatical constraints of gender exist.

7. ACKNOWLEDGEMENT

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

8. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, no. 5217, pp. 1632–1634, 1995.

[3] V. van Wassenhove, K. W. Grant., and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proceedings of the National Academy of Sciences*, vol. 102, no. 4, pp. 1181–1186, 2005.

[4] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 687–693.

[5] M. Fleischman and D. Roy, "Grounded language modeling for automatic speech recognition of sports video," in *ACL-08 HLT*, 2008, pp. 121–129.

[6] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5020–5024.

[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, Sept 2003.

[8] J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5774–5778.

[10] Y. Moriya and G. J. F. Jones, "Lstm language model adaptation with images and titles for multimedia automatic speech recognition," in *Spoken Language Technology Workshop (SLT)*, 2017.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 55–59.

[13] Y. Miao, L. Jiang, H. Zhang, and F. Metze, "Improvements to speaker adaptive training of deep neural networks," in *Spoken Language Technology Workshop (SLT)*, 2014, pp. 165–170.

[14] M. Rouvier and B. Favre, "Speaker adaptation of DNN-based ASR with i-vectors: Does it actually adapt models to speakers?," in *Interspeech*, 2014, pp. 3007–3011.

[15] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017.

[16] B. Herve and G. Gregory, "Improving speaker diarization of tv series using talking-face detection and clustering," in *ACM Multimedia (ACMM)*, 2016, pp. 157–161.

[17] E. Vanmassenhove, C. Hardmeier, and A. Way, "Getting gender right in neural machine translation," in *Empirical Methods in Natural Language Processing (EMNLP)*. 2018, pp. 3003–3008, Association for Computational Linguistics.

[18] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[20] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech*, 2015, pp. 2–6.

[21] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, 2010, pp. 1045–1048.

[22] S Hochreiter and J Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–80, 1997.

[23] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, pp. 310–318, 1995.

[26] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, 2002.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 1–4.