Speaker Adaptive Audio-Visual Fusion for the Open-Vocabulary Section of AVICAR

Leda Sari[1,2], Mark Hasegawa-Johnson[1,2], Kumaran S[3], Georg Stemmer[3], Krishnakumar N. Nair[3]

[1]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, US
[2]Beckman Institute, University of Illinois Urbana-Champaign, US
[3]Intel Corporation

{lsari2,jhasegaw}@illinois.edu, {kumaran.s,georg.stemmer,krishnakumar.n.nair}@intel.com

Abstract

This experimental study establishes the first audio-visual speech recognition baseline for the TIMIT sentence portion of the newly re-aligned AVICAR dataset which is recorded in a real, noisy car environment. We use an automatic speech recognizer trained on a larger dataset to generate audio-only recognition baseline for AVICAR. We utilize the forced alignment of the audio modality of AVICAR to get training targets for the convolutional neural network based visual front end. Based on our observation that there is a great amount of variation between visual features of different speakers, we apply feature space maximum likelihood linear regression (fMMLR) based speaker adaptation to the visual features. We use audio modality either as additional features or use Gaussian posteriors of audio in order to estimate speaker dependent transformation matrices. In addition, we rescore the audio-only lattices using visual unit posteriors obtained by classifying speaker adapted features to generate the word hypotheses. We report the first audio-visual results for TIMIT subset of AVICAR and show that the word error rate of the proposed audio-visual system is significantly better than that of the audio-only system.

Index Terms: audio-visual speech recognition, neural networks, speaker adaptation

1. Introduction

The aim of audio-visual speech recognition (AVSR) is to exploit visual data such as mouth movements of the speaker to improve the accuracy of audio-only automatic speech recognition (ASR) especially in noisy acoustic conditions.

Early AVSR systems are mainly based on multi-stream hidden Markov models (HMMs) [1, 2, 3]. Recently, neural network based models are used for AVSR and lipreading in order to extract features or the networks are used as classifiers. For example, in [4, 5], deep belief networks and convolutional neural networks (CNNs) are used as feature extractors. In [6], neural networks are used as phoneme classifier given audio-visual input.

This experimental study establishes the first baseline results for the TIMIT sentence portion of the AVICAR dataset. The paper shows how the audio component can guide learning the complementary information in video modality in three ways. First, we use the forced alignments of the audio to get training labels for feature extraction of video component which is achieved by CNNs. Since visual realization of speech units vary among speakers, we apply speaker adaptation to the visual features using feature space maximum likelihood linear regression (fMLLR) [10]. In order to estimate the adaptation model, we either concatenate audio features to CNN-based visual features or use the alignment statistics from audio modality to learn a transform for video in a cross-modal setting. Finally, we use the visual unit posteriors from neural networks trained on visual features to rescore audio-only recognition lattices and to get the word recognition hypotheses.

In AVSR applications, training targets labels are either provided with the dataset such as time intervals of spoken digits in digit recognition studies or the AVSR dataset itself is used to train an ASR and get the phonetic alignment [16]. In lipreading literature whose aim is to transcribe lip movements without the use of audio, the 'Lip reading in the wild' dataset [7] uses the idea of using an external recognizer trained on large amounts of data to check their labels during dataset construction. In our study, we also use an external dataset but to obtain our baseline ASR results by decoding the AVICAR dataset. In this study, we do not only use audio for data labeling but also for multimodal speaker adaptation.

There are studies that use audio alignment information to guide feature normalization or adaptation using linear discriminant analysis (LDA) and fMLLR [8, 4]. However, these studies use discrete cosine transform [8] or deep belief network [4] based visual features. In this work, we apply audio-guided transformation to CNN-based features. In [9], fMLLR is used for adaption of shape and appearance features for lipreading task but their transformation is unimodal as opposed to cross-modal training for AVSR in this work. In [5], speaker variability is handled by training speaker specific networks rather than speaker-independent networks. However, this is only feasible for small number of speakers with sufficient amount of adaptation data. In our study, we use speaker-independent networks and we do not require transcriptions for the test set speakers.

Our rescoring mechanism resembles the weighted combination of log-likelihoods of audio and video modalities in the multi-stream HMM based studies [4, 5, 1, 3]. However, we only scale the visual component and our probabilities are obtained from the softmax layer of neural networks rather than HMMs.

The rest of the paper is structured as follows. The ways by which audio component is utilized to extract the complementary information in video is described in Section 2. Experimental setups and results are presented in Section 3 and the paper is concluded in Section 4.

2. Audio-Visual Fusion

In our AVSR system, audio component is utilized in three ways:

1. State-level forced alignments are converted into visual unit alignments and used as output targets for the CNN used as visual feature extractor.

2. Audio features and Gaussian posteriors obtained from audio-only GMM-HMM system are used in fMLLR-based [10] speaker adaptation of AV or visual features.

3. Audio lattices rescored by the viseme probabilities computed from visual data are used to generate the word hypotheses.

The following subsections will present these points.

2.1. Visual Feature Extraction

In earlier AVSR studies, transform based features [1], shape and appearance models are used in order to extract visual features. With the recent advances in neural network based systems, hidden layer activations of the networks started to replace those features in AVSR [4, 5]. In this study, we use CNNs as our visual feature extractors.

In the visual front end, we first determine the mouth area of each video frame based on the facial landmarks, then crop the mouth area into a fixed sized window. Cropped images are converted to grayscale. For each frame, neighboring frames are used as context windows and fed into convolutional layers. These layers are followed by fully connected layers and a softmax classification layer.

Classification targets are obtained using the audio modality. Forced alignment of the training data is used to get frame-level phonetic labeling of the data. In our systems we used two types of targets, visemes and clustered visual units. To determine the viseme targets we applied the widely used phoneme to viseme mapping of [11], which has 14 visemes, to the phonetic labels. Since there is not a general agreement on the viseme classes [12], we also explore data-driven visual units as an alternative.

In order to get the data-driven, clustered visual units, we initialize our cluster centers as the mean visual vector associated with each phone based on the phonetic alignment of the data. We compute the top K closest vectors using Euclidean distance and merge two units if they are mutually within their K-neighborhoods. After combination, we recompute the mean vectors for the new set of clusters and continue until the desired number of units are obtained. This method can be interpreted as a modified version of the shared k-nearest neighbor which does not have the update step [13]. As we start merging with phonemes, we retain the relationship between phonemes and the visual units. Thus we have a phoneme to visual unit map.

Once we train CNNs using either set of targets, the visual features are computed from the last convolutional layer of the CNN.

2.2. fMLLR-based Speaker Adaptation

Although the hidden layer activations of CNNs are robust to variations such as shift, they are not invariant to speakers. Previous studies also show that lip features are highly speaker-dependent [9]. In order to achieve speaker normalization of the CNN-based visual features, we use the fMLLR technique which is widely used method for speaker adaptation in audio-only ASR. In this technique, features are modified by an affine transform where the transformation matrix is learned using the expectation-maximization algorithm [10]. Estimation of the speaker specific transformation matrices require the Gaussian posteriors from the GMM-HMM along with the feature vectors. Depending on how the audio information is used we have two ways of estimating transformations.

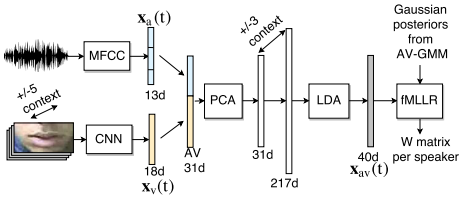

Figure 1: *Flowchart for early fusion*

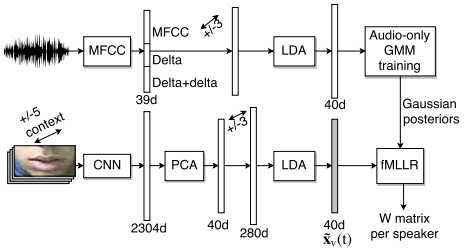

Figure 2: *Flowchart for implicit fusion*

In the first method, which we call early fusion, both the features and the Gaussian posteriors are audio-visual. As shown in Fig. 1, MFCC features of audio ($\mathbf{x}_a(t)$) are concatenated to the activations obtained from the last convolutional layer of the CNN ($\mathbf{x}_v(t)$). Then, we apply mean and variance normalization and principal component analysis (PCA). To make use of the temporal context, we concatenate neighboring frames and reduce the dimension by linear discriminant analysis (LDA). If the AV features after LDA are denoted by $\mathbf{x}_{av}(t)$, then the fMLLR transformed observations $\hat{\mathbf{o}}(t)$ at time t can be written using the augmented AV features $\zeta_{av}(t) = [\mathbf{x}_{av}(t);\ 1]$ as $\hat{\mathbf{o}}(t) = \mathbf{W}\zeta_{av}(t)$. The i-th row of the transformation matrix $\mathbf{W}$ is found by $\mathbf{w}_i = (\alpha\mathbf{p}_i + \mathbf{k}_i)\mathbf{G}_i^{-1}$ where $\mathbf{p}_i$, α, and $\mathbf{k}_i$ can be calculated as shown in [10] and

$$\mathbf{G}_i = \sum_m \frac{1}{\sigma_i^{(m)2}} \sum_t \gamma_{av}^{(m)}(t)\zeta_{av}(t)\zeta_{av}(t)^T. \qquad (1)$$

In (1), superscript T denotes transposition, $\sigma_i^{(m)}$ is the i-th diagonal element of the covariance matrix and $\gamma_{av}^{(m)}$ is the posterior probability of the m-th Gaussian trained on AV features.

In the second method, which we call implicit fusion, visual features are used but the Gaussian posteriors are obtained by the audio-only GMM-HMM system. As shown in Fig. 2, audio and video components are processed separately and fusion occurs while estimating the fMLLR transformation matrices. If PCA and LDA applied visual features are denoted by $\tilde{\mathbf{x}}_v(t)$, and the Gaussian posteriors obtained from the audio is denoted by $\gamma_a^{(m)}(t)$, estimation of $\mathbf{G}_i$ matrices can be written as

$$\mathbf{G}_i' = \sum_m \frac{1}{\sigma_i^{(m)2}} \sum_t \gamma_a^{(m)}(t)\zeta_v(t)\zeta_v(t)^T \qquad (2)$$

where $\zeta_v(t) = [\tilde{\mathbf{x}}_v(t);\ 1]$.

Once we get the speaker adapted audio-visual or visual features using these methods, we train a fully-connected network for visual unit classification and generate visual unit posteriors to be used in the decision fusion stage which is rescoring of the audio-only lattices using visual unit posteriors.

2.3. Generating Word Hypotheses

In our AVSR system, word hypotheses are generated from rescored audio-only recognition lattices. Rescoring is done by

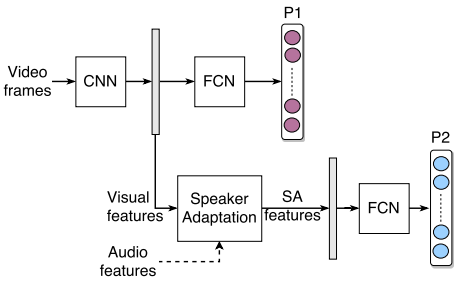

Figure 3: *Steps at which visual posteriors (P1 and P2) are generated, FCN denotes fully-connected network*

adding weighted log posterior probabilities of visual units to the log probability of the corresponding state in the lattice. We have the correspondence between states and visemes since there is a mapping from states to phonemes and phonemes to visemes. In the case of clustering-based units, we start clustering based on the phonetic identities of the frames, therefore we retain a many-to-one mapping from phonemes to visual units and thus from states to visual units.

The visual unit probabilities are computed from the softmax layer of a neural network which is either a CNN trained on video frame pixels (P1 in Fig.), i.e. the feature extractor network, or a fully connected network trained on speaker adapted features (P2 in Fig.).

The combination weight can be constant for all noise conditions in the dataset or they can be chosen inversely proportional to the estimated SNR levels of the audio for each recording condition. SNR estimates are obtained by the average ratio of the energy in the longest speech portion of training utterances to the energy in the silence portion of the utterances.

3. Experimental Results and Discussion

We performed our AVSR experiments on the read TIMIT sentences of AVICAR12 [14], the synchronized version of the AVICAR corpus described in [15]. The dataset consists of digit sequences and TIMIT sentences read in a car environment. Our training set contains 3.48hr speech from 61 speakers and our test set contains 1.14hr speech from 21 speakers who are distinct from the training speakers. There are five recording conditions with different noise levels. They depend on the speed of the car (35/55mph), windows being open or closed (D/U) and idling car engine (IDL). Prior work on AVICAR has published word error rates (WERs) only for isolated digits and digit sequences (audio-only [16], 19.26%, visual-only [17] 37.87%, and audiovisual [8]). Table 1 compares the audio-only ASR results of [16] which is trained only on the IDL condition and the ASR baseline of this study which is trained on a larger external dataset. Except IDL and 55D conditions, our ASR has lower WER. Worse performance in 55D results from the fact that we work on TIMIT sentences portion with a much larger vocabulary as compared to the digit strings with limited vocabulary. The only published audiovisual WER [8] is given in Table 2 and is more than twice as high as any WER reported in this paper. However, these results are reported only for the telephone digit string recognition task. However, our aim is to use AVSR on the more challenging portion of the AVICAR dataset which contains read sentences from the TIMIT corpus [18].

The baseline audio-only ASR system is trained on a larger external database than AVICAR using a deep neural network based setup in Kaldi with 3897 triphone HMM states, and using a unigram language model. Speaker adaptation and WER scoring are performed using Kaldi [19].

Table 1: *Comparison of the audio baselines from [16] and this study for different noise conditions in the AVICAR corpus*

	Avg	35U	55U	35D	55D	IDL
ASR baseline [16]	19.26	13.16	21.40	24.23	34.95	4.22
ASR baseline	22.73	12.89	14.89	19.59	56.78	5.94

Table 2: *Relative improvement (in %) in WER between audio-only ASR and AVSR in [8] that recognizes digit sequences*

	IDL	35U	55U	35D	55D	Avg
ASR [8]	35	51	57	62	77	56
AVSR [8]	35	46	51	52	63	49
Change (%)	0	9.8	10.5	16.1	18.2	12.5

In the following subsections, details of the experimental AVSR setups and the results will be presented.

3.1. Visual Unit Classification

In order to get training labels for the visual feature extractor, we convert the state-level alignment into phonemes and then to 14 viseme classes using the map of [11] or by shared k-nearest neighbor clustering. The number of clusters is chosen to be 22 which results in the lowest test WER in the AVSR system. The feature extractor CNN applies rectified linear unit nonlinearity on the results of convolutions which is then followed by local response normalization and max pooling. The inputs of the networks are cropped mouth areas determined by Dlib facial landmark detectors. To capture the mouth movements and to get contextual information, ± 5 neighboring frames are used as additional input channels. If the kernel size, number of output channels and the pooling factor are represented as a triplet, the CNN layers are summarized as (5,48,2), (3,128,2), (3,192,1), (3,192,2), (5,128,2). Since we start with fixed-sized cropped images of size 96x48, at the last convolutional layer we have 128 channels each of size 6x3. To generate visual features for fusion, we concatenate these activations and get 2304-d vectors. The softmax output layer has 14 or 22 units depending on the training label type which are visemes or clustered visual units, respectively. The networks are trained using Tensorflow [20] to minimize cross-entropy.

In order to adapt the 2304-d visual features to speakers, we performed 5 iterations of expectation-maximization for fMLLR transform estimation. Once the speaker specific matrices are estimated, adapted features are fed into fully-connected networks. The fully-connected networks used for classifying speaker adapted features and obtaining visual unit posteriors for lattice rescoring have 3 hidden layers with 128, 128 and 32 nodes followed by a softmax layer where each node represents a visual unit. Depending on the visual unit type, the number of nodes is 14 or 22. As for the CNNs, the networks are trained to minimize the cross-entropy measure.

Table 3 summarizes the training and test set accuracy of the visual unit classifiers, for both the visual feature extractor and the fully connected network trained on the speaker adapted features. The results are reported for the cases where we have viseme or clustering-based visual unit (CVU) training

Table 3: *Visual unit classification accuracy (in %) depending on the visual unit target, viseme or clustering-based visual units (CVU), used depending on the input features of the networks which are unadapted or speaker adapted visual features*

		Train	Test
Feature extraction	Viseme	52.85	47.41
	CVU	45.43	40.95
Speaker adaptation	Early fusion	72.38	67.44
	Viseme	55.38	51.09
	CVU	53.54	49.55

targets. Speaker adaptation strategies used in these results are both early and implicit speaker adaptation strategy described in Section 2.2. For early fusion, only viseme based results are reported as the final WER of the AVSR system is worse than using unadapted visual features although the visual unit classification accuracy is higher. Better classification performance of speaker adaptation with early fusion strategy results from the fact that the speaker adapted input features of the fully connected network contain audio features along with the visual ones.

3.2. Lattice Rescoring and the WER

The final WER of the AVSR systems are obtained by audio-lattice rescoring as described in Section 2.3. The log-posterior combination weight λ is chosen by grid search. For the variable weighting scheme, λ correspond to the scaling factor for the multiplicative inverse of the SNR levels. In both cases, the optimal λ is chosen such that we achieve the lowest WER for the training data as we do not have a separate held-out set.

Table 4 shows the WER of the audio only baseline, WER of the rescored lattices with visual unit posteriors obtained from visual features for the test data. The first set of results show the WER obtained by rescoring with the posteriors obtained from the feature extractor CNN when training targets are visemes or clustering-based units. The second set of results show the WER for rescoring with posteriors obtained from speaker adapted visual features either by early fusion for visemes or by implicit fusion for both types of visual units. Early fusion with clustering-based units is not included as early fusion has worse performance than using unadapted features.

Adding visual unit posteriors from unadapted visual features reduces the WER to 21.12 and 21.15% with viseme targets and clustering-based units, respectively, as compared to the audio-only baseline of 22.73%. Using posteriors from speaker adapted features reduces the WER further if we apply implicit fusion. We achieve similar performance improvement with clustering-based units as with the visemes. The lowest WER 20.91% is achieved when we use CVU targets and the implicit fusion for speaker adaptation. The relative reduction in WER is 8% in this setup. This result suggests that audio alignment helps adaptation of the video component and allows us to extract additional information from video especially at the lattice rescoring phase. As mentioned earlier, WER of the speaker adapted system with concatenated features (early fusion) results in higher WER than unadapted features which suggests that the fMLLR estimation with concatenated audio-visual information is not reliable. These observations lead to the conclusion that audio is better aligned with the speech content and audio can guide learning complementary information from the video input. Another observation is that data-driven clustering based units achieve comparable performance to the hand-crafted and tabulated viseme classes.

Table 4: *WER (in %) after late fusion of posteriors obtained from video or ISAF classifiers depending on the VU of the feature extractor. For comparison, audio-only WER is 22.73%.*

		WER (%)
Audio	-	22.73
Feature extractor	Visemes	21.12
	CVU	21.15
Speaker Adaptation	Early	21.26
	Visemes	21.02
	CVU	20.91

Table 5: *Comparison of the environment dependent WER for audio-only ASR system and the best performing AVSR system*

	Avg	35U	55U	35D	55D	IDL
ASR	22.73	12.89	14.89	19.59	56.78	5.94
AVSR	20.91	13.03	14.78	15.77	51.98	5.79

If we apply a t-test with the null hypothesis that there is no WER difference between two setups, we see that both the late fusion of posteriors from video frames and speaker adapted feature based posteriors perform significantly better than the audio-only baseline at the level of p=0.001. However, the difference between two AVSR methods is not significant.

If we compare the noise condition dependent WER from audio-only baseline with the best performing setup, we obtain the results in Table 5. We observe that we achieve WER reduction in almost all noise conditions and the largest absolute improvement is in the noisiest condition 55D which is the condition for which the use of complementary visual information is crucial.

4. Conclusions

In this study, we showed that how audio component can be used to guide extracting information from the video component in AVSR task in noisy car environment. First, forced state-level alignment of the audio allows us to determine the visual unit targets for the feature extractor CNN. Secondly, we used Gaussian alignments of the audio component to estimate fMLLR matrices for speaker adaptation of the visual features as we observed that the CNN-based features are not invariant to speakers. Thirdly, we used rescored audio-only recognition lattices with visual unit posteriors to generate our word hypotheses from our AVSR system. The visual units are either hand-crafted visemes or data-driven clustering based units and the posteriors are obtained by classifying speaker adapted visual features into these units. Experiments on the TIMIT sentence section of AVICAR corpus establishes AVSR baseline for the dataset and demonstrates that the ASR performance is improved by 8% using the proposed speaker adaptation strategy. We also observed that the largest gains are obtained for the noisiest conditions and the use of data-driven clustering-based units achieve similar performance as the viseme based setup.

5. Acknowledgements

This work is supported by Intel. We would like to thank Josef Bauer from Intel for generating the ASR lattices for the audio-only experiment.

6. References

[1] 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, 2003.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[3] J. Huang, E. Marcheret, and K. Visweswariah, "Rapid feature space speaker adaptation for multi-stream HMM-based audio-visual speech recognition," in *Proc. ICME*, pp. 338–341.

[4] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7596–7599.

[5] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[6] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. ICASSP*. IEEE, 2015, pp. 2130–2134.

[7] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.

[8] R. Navarathna, D. B. Dean, P. J. Lucey, S. Sridharan, and C. B. Fookes, "Recognising audio-visual speech in vehicles using the AVICAR database," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, 2010, pp. 110–113.

[9] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2722–2726.

[10] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] S. Lee and D. Yook, "Audio-to-visual conversion using hidden Markov models," *PRICAI 2002: Trends in Artificial Intelligence*, pp. 9–20, 2002.

[12] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *Proc. 19th EUSIPCO*. IEEE, 2011, pp. 2109–2113.

[13] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," *IEEE Transactions on computers*, vol. 100, no. 11, pp. 1025–1034, 1973.

[14] S. S. T. Group, "AVICAR corpus, version AVICAR12," March 2013, http://www.isle.illinois.edu/sst/AVICAR/.

[15] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang, "AVICAR: Audio-visual speech corpus in a car environment." in *Interspeech*, 2004, pp. 2489–2492.

[16] B. Lee, "Robust speech recognition in a car using a microphone array," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2006.

[17] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang, "Lipreading by locality discriminant graph," in *Proc. ICIP*, vol. III, 2007, pp. 325–8.

[18] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[19] 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 *Proc. ASRU*, Dec. 2011.

[20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.