Improving Lip-reading with Feature Space Transforms for Multi-Stream Audio-Visual Speech Recognition

Jing Huang and Karthik Visweswariah

IBM T.J. Watson Research Center
Yorktown Heights, NY, USA
{jghg,kv1}us.ibm.com

10.21437/Interspeech.2005-373

Abstract

In this paper we investigate feature space transforms to improve lip-reading performance for multi-stream HMM based audio-visual speech recognition (AVSR). The feature space transforms include non-linear Gaussianization transform and feature space maximum likelihood linear regression (fMLLR). We apply Gaussianization at the various stages of visual front-end. The results show that Gaussianizing the final visual features achieves the best performance: 8% gain on lip-reading and 14% gain on AVSR. We also compare performance of speaker-based Gaussianization and global Gaussianization. Without fMLLR adaptation, speaker-based Gaussianization improves more on lip-reading and multi-stream AVSR performance. However, with fMLLR adaptation, global Gaussianization shows better results, and achieves 18% over baseline fMLLR adaptation for AVSR.

1. Introduction

Recently audio-visual speech recognition (AVSR) has attracted significant interest as a means of improving performance and noise robustness over audio-only speech recognition (ASR) [1, 2, 3]. However, most AVSR research work in the literature, has concentrated on databases recorded under ideal visual conditions with subjects' frontal face, very limited variation in head pose, uniform lighting, and constant background. This kind of visual data is often unrealistic in practical scenarios, where the subject's posture and the environment lighting are hard to control, for example, automobiles, offices, and trading floors.

In such real-life applications, robust extraction of the visual speech information becomes challenging problem, since it requires accurate tracking of the speaker's face and facial features (e.g., mouth corners, possibly lip contours), as well as successful compensation for head pose and lighting variations in the final visual speech features. Preliminary results by [4] show that lip-reading (i.e. visual-only recognition) degrades dramatically in visually challenging domains (offices, automobiles). In addition AVSR performance lags significantly compared to the performance achieved on data recorded in studio-like environment, as measured by visual-only *word error rate* (WER) and relative WER reduction in bimodal vs. audio-only ASR in noise.

These facts motivate us to improve lip-reading by visual feature transformation and adaptation to realistic environments. In the framework of multi-stream HMM-based audio-visual speech recognition, both audio and visual HMM output distributions are modeled with mixtures of diagonal covariance Gaussians. We examine the visual data, both before linear discriminant analysis (LDA) and after LDA, we notice the data distribition are not so gaussian (see Figure 1). Therefore we naturally apply feature space Gaussianization technique [5] to the visual data to better fit to the Gaussian modeling assumption. Another advantage of Gaussianization is that both test and *training* speakers are warped to the same space which naturally leads to a form of speaker adaptive training (SAT) through non-linear transforms. In addition to the speaker-based feature space Gaussianization in [5], we also study the effectiveness of global Gaussianization transform [6] on the visual data, because in run-time we may not have enough data to estimate cumulative distribution function (CDF) for a speaker. In this case, the global Gaussianization transform is handy.

Speaker adaptation is a key successful technique that is used in most of the stat-of-the-art ASR systems. In [7] fast feature space speaker adaptation is investigated for multi-stream audio-visual speech recognition, and feature-space maximum likelihood linear regression (fMLLR, also known as constrained MLLR) [8] is shown to be very effective (achieving almost 60% relative gain for multi-stream AVSR in the noisy case). In this paper, we again add the fMLLR adaptation on top of Gaussianization technique and indeed we obtain additive gain.

We want to make a note here that although we only focus on feature space transforms (such as non-linear Gaussianization and fMLLR) to improve visual-only lip-reading in this paper, these techniques are also quite effective for audio data [5] and multi-stream AVSR [7]. We also present results of improved lip-reading in the multi-stream AVSR framework in Section 4.

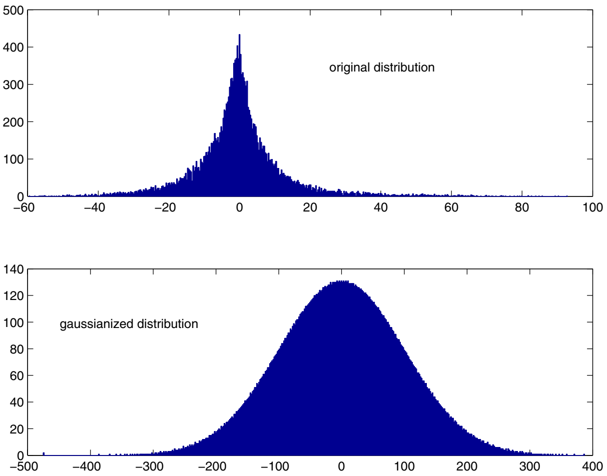

Figure 1: *One example of the original distribution of visual data at dimension 3 and its gaussianized distribution.*

The paper is structured as follows: The visual processing steps for multi-stream AVSR is discussed in Section 2. Section 3 briefly describes speaker-based Gaussianization and global Gaussianization. Experimental setup and results are reported in Section 4, and conclusions are drawn in Section 5.

2. Visual Features and the Multi-stream AVSR System

Our experiments in this paper are conducted on the audio-visual database collected with the IBM infrared headset [9]. The infrared headset is specially designed equipment that captures the video of the speaker's mouth region, independently of the speaker's movement and head pose. It reduces environmental lighting effect on captured images, allowing good visibility of the mouth ROI even in a dark room. Since the headset consistently focuses on the desired mouth region, face tracking is no longer required. Eliminating this step improves the visual front end robustness and reduces CPU requirements by approximately 40% [11].

Visual features are extracted from the region of interest (ROI). We first estimate the location of the ROI, which contains the area around the speaker's mouth (see [9] for details). Following ROI extraction, the visual features are computed by applying a two-dimensional separable DCT to the sub-image defined by the ROI, and retaining the top 100 coefficients with respect to energy. The resulting vectors then go though a pipeline consisting of intra-frame linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT), temporal interpolation, and feature mean normalization, producing a 30-dimensional feature stream at 100Hz. To account for inter-frame dynamics, fifteen consecutive frames in the stream are joined and subject to another LDA/MLLT step to give the final visual feature vectors with 41 dimensions.

Notice that we could gaussianize the final 41-dim visual features, or gaussianize the initial 100-dim DCT coefficients, or gaussianize at the each step before/after two steps of LDAs.

In parallel to the visual feature extraction, audio features are also obtained, time synchronously, at 100 Hz. First, 24 mel frequency cepstral coefficients of the speech signal are computed over a sliding window of 25 msec, and are mean normalized to provide static features. Then, nine consecutive such frames are concatenated and projected by means of LDA/MLLT onto a 60-dimensional space, producing dynamic audio features.

In the multi-stream HMM based decision fusion approach, the single-modality observations are assumed generated by audio-only and visual-only HMMs of identical topologies with class-conditional emission probabilities $P_a(\mathbf{o}_{a,t})$ and $P_v(\mathbf{o}_{v,t})$, respectively. Both are modeled as mixtures of Gaussian densities. Based on the assumption that audio and visual streams are independent, we compute the joint probability $P_{av}(\mathbf{o}_{av,t})$ as follows [10]:

$$P_{av}(\mathbf{o}_{av,t}) = P_a(\mathbf{o}_{a,t})^{\lambda} \times P_v(\mathbf{o}_{v,t})^{1-\lambda}$$

Exponent λ is used to appropriately weigh the contribution of each stream, depending on the "relative confidence" on each modality. Exponents can be fixed or time dependent [12]. We chose fixed weights in this paper.

3. Feature Space Gaussianization

One obvious advantage of feature space Gaussianization is that it is purely data-driven and completely unsupervised — no need for a first-pass decoding step for adaptation. We follow the simple and fast approach in [5]: Gaussianization transform is implemented as a simple lookup table where the entries are given by the inverse Gaussian CDF sampled uniformly in $[0, 1]$. We used one million samples for our experiments. For each speaker we first sort all the samples on each dimension. Then compute the empirical CDF by equation

$$F_0(x_i) = \frac{rank(x_i)}{N}$$

where $rank(x_i)$ is the rank of x_i in the sorted list of samples on dimension i. This value is used to locate the lookup table and get back the transformed feature y_i for x_i.

Instead of computing the empirical CDF $F_0(x_i)$ from data of each speaker, Chen and Gopinath [6] propose to use a mixture of Gaussian CDF's estimated from training data. The parametric approach allows us to estimate global Gaussianization transform for each dimension and apply to test data directly without the need for a reliable estimate of the empirical CDF.

4. Experiments and Results

4.1. Experimental Setup

As mentioned before our experiments are evaluated on the audio-visual database collected with the IBM infrared headset. The AVSR system is built on 22kHz audio and 720x480 pixel resolution at 30 Hz video. A total of 107 subjects uttering approximately 35 random length *connected digit* sequences. We split 107 speakers in our infrared headset data into training set and testing set: 87 speakers are used for training, and the remaining 20 speakers are used for testing, and there is no overlap speakers in training and testing sets. The training data has about 4 hours of speech, and the test data has around 1 hour speech. Both training and testing data have an average SNR of 20dB. In addition to this clean test data which matches the training data, another noisy test set is built by artificially corrupting the test set with additive "speech babble" noise resulting in an average SNR of 7dB. Recognition results are presented on both clean and noisy test sets.

The recognition system uses three-state, left-to-right phonetic HMMs with 159 context-dependent states (the context is cross-word, spanning up to 5 phones to either side) and $2,600$ Gaussian mixture components with diagonal covariances.

We emphasize here again that in this paper we focus on evaluating different feature space transforms for visual data to improve the performance of multi-stream AVSR. Therefore we leave the baseline audio system alone and compare different feature space transforms on performance of visual-only and multi-stream audio-visual (AV) recognition. Adaptation with fMLLR is also done on visual channel only (see [7] for results of multi-stream fMLLR).

4.2. Results

The results are presented as word error rate (WER) for visual-only (V), and multi-stream audio-visual (AV) recognition. Baseline audio-only results in clean and noisy condition are also listed as comparision to the AV results. These recognition results are run by an IBM Viterbi decoder (therefore the baseline numbers are not the same as those in [7] which uses the IBM stack decoder).

In Table 1, we compare the results of Gaussianization at different stage of the visual front-end:

- G1: Gaussianize just the initial 100-dimensional features.
- G2: Gaussianize at every step of the visual front-end: starting with 100-dim features, Gaussianize them and train LDA+MLLT, then Gaussianize 30-dim features and train LDA+MLLT, finally Gaussianize the final 41-dim features.
- G3: Gaussianize just the final 41-dimensional features.
- G4: all the above are speaker-based Gaussianization, this one uses one global Gaussianization transform.

	Clean			Noisy		
	A	V	AV	A	V	AV
baseline	2.1	36.0	1.4	15.0	36.0	8.8
G1	-	34.3	1.3	-	34.3	8.3
G2	-	33.8	1.2	-	33.8	8.5
G3	-	33.2	1.2	-	33.2	8.2
G4	-	35.2	1.3	-	35.2	9.0

Table 1: *Comparison of Gaussianization on different visual features.*

It is surprising that G3 is the best: just Gaussianizing the final 41-dimensional features is better than Gaussianizing features at every step. It is also clear that speaker-based Gaussianization is better than global Gaussianization: visual-only WER for G3 is 33.2%, improves 8% relative to the baseline 36.0%; visual-only WER for G4 is 35.2%, improves only a little on the baseline.

The fourth and the seventh columns are results of multi-stream AVSR. The improvement on the visual recognition carries over to AVSR: using G3 AV improves in the clean condition from baseline 1.4% to 1.2%, 14% relative gain; in the noisy condition from baseline 8.8% to 8.2%, 7% relative gain. Using G4 improves a little in the clean condition, but degrades a little in the noisy condition. We would suggest that in test time, use global Gaussianization first. When there is enough test data, then switch to speaker-based Gaussianization for better performance.

The above Gaussianization transforms are applied to each training and test speaker. fMLLR adaptation follows the Gaussianization of test speakers. Our fMLLR adaptation is unsupervised: for each speaker, we use the baseline speaker independent audio and visual models to get initial multi-stream AVSR decoding transcripts, and use these transcripts to compute fMLLR transform for visual stream. Then the transformed testing features are used to get the final adapted results. We keep the stream weights fixed, 0.7 for audio stream, and 0.3 for video stream.

We observe these interesting facts from Table 2: whether to use 1.4%-WER transcripts from AVSR on clean data or to use 8.8%-WER transcripts from AVSR on noisy data, fMLLR adapted visual WERs are about

	Clean		Noisy	
	V	AV	V	AV
baseline	36.0	1.4	36.0	8.8
baseline+fMLLR	22.4	1.1	22.9	7.3
G3+fMLLR	20.0	1.0	20.7	7.0
G4+fMLLR	20.8	0.9	21.4	7.3

Table 2: *Comparison of Gaussianization followed by fMLLR adaptation.*

the same: 22.4% from clean AVSR transcripts and 22.9% from noisy AVSR transcripts. Secondly, G4 is worse than G3 for visual-only and AVSR (see Table 1); however, when combined with fMLLR, G4+fMLLR is better than G3+fMLLR for AVSR on clean test data: G4+fMLLR improves baseline+fMLLR from 1.1% to 0.9%, 18% relative gain; while G3+fMLLR improves baseline+fMLLR from 1.1% to 1.0%.

On the visual-only performance, G3+fMLLR is still better than G4+fMLLR: on clean test data, G3+fMLLR improves baseline+fMLLR from 22.4% to 20.0%, 11% relative gain; G4+fMLLR improves baseline+fMLLR from 22.4% to 20.8%, 7% relative gain; on noisy test data, G4+fMLLR shows even less gain over baseline+fMLLR than its results on the clean test data.

However, both G3+fMLLR and G4+fMLLR do not improve much over baseline+fMLLR AVSR result on the noisy test data. This suggests that there may be a limitation of contribution from visual channel on the noisy audio data. Gaussianization step does not help much and fMLLR alone is sufficient for adaptation.

5. Conclusion

In summary, we investigate various feature space transforms to improve lip-reading performance for multi-stream HMM based audio-visual speech recognition. These transforms include non-linear Gaussianization transform and feature space maximum likelihood linear regression (fMLLR). We apply feature space Gaussianization at the various stages of visual front-end. The results show that Gaussianizing the final visual features is sufficient and achieves the best performance. We also compare the performance of speaker-based Gaussianization and global Gaussianization. Without fMLLR adaptation, speaker-based Gaussianization improves more on lip-reading performance and multi-stream AVSR performance. However, with fMLLR adaptation, global Gaussianization shows better results. This is beneficial since global Gaussianization does not need run-time estimation of empirical CDF as speaker-based Gaussianization does.

It is disappointing that the improved lip-reading does not contribute well to noisy speech recognition, even when the WER improves from 36% to 20% (44% relative gain) after fMLLR adaptation. In this case, fMLLR on the audio stream is essential, and the multi-stream fMLLR which combines Gaussian posterior counts from both streams is much more effective than audio fMLLR alone (see [7]).

6. References

[1] Janin, A., Ellis, D., and Morgan, N., “Multi-stream speech recognition: Ready for prime time?”, *Proc. Europ. Conf. Speech Technol.*, pp. 591–594, 1999.

[2] Dupont, S. and Luettin, J., “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, 2(3): 141–151, 2000.

[3] Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A.W., “Recent advances in the automatic recognition of audio-visual speech,” *Proc. IEEE*, 91(9): 1306–1326, 2003.

[4] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” *Europ. Conf. Speech Commun. Technol.*, 2003.

[5] G. Saon, S. Dharanipragada and D. Povey, “Feature Space Gaussianization,” *Proc. ICASSP*, 2004.

[6] S. Chen and R. Gopinath, ”Gaussianization,” *Proc. NIPS’00*, Denver, 2000.

[7] J. Huang, E. Marcheret, K. Visweswariah, “Rapid Feature Space Speaker Adaptation for Multi-Stream HMM-based Audio-Visual Speech Recognition”, *IEEE Int. Conf. on Multimedia & Expo*, 2005, to appear.

[8] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Technical report, TR 291, Cambridge University*, 1997.

[9] J. Huang, G. Potamianos, J. Connell and C. Neti, “Audio-Visual Speech Recognition Using an Infrared Headset,” *Speech Communication*, Dec. 2004.

[10] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” IEEE Trans. Multimedia, 2(3):141–151, 2000.

[11] J. Connell, N. Haas, E. Marcheret, C. Neti, G. Potamianos, S. Velipasalar, “A Real-Time Prototype for Small-Vocabulary Audio-Visual ASR,” *IEEE Int. Conf. on Multimedia & Expo*, 2003.

[12] A. Garg, G. Potamianos, C. Neti, T. Huang, “Frame-Dependent Multi-Stream Reliability Indicators for Audio-Visual Speech Recognition,” *Int. Conf. Acoustic Speech and Signal Processing*, 2003.