自動リップリーディングのための位置正規化

Oscar Vanegas, 徳田恵一, 北村正
名古屋工業大学

Location Normalization for Automatic Lipreading

Oscar Vanegas, Keiichi Tokuda, Tadashi Kitamura
Nagoya Institute of Technology

1 Introduction

Some methods using images from the lips in motion are proposed for audio-visual speech recognition. However problems arise when a large variation of lip location is present on images. Therefore, this paper presents a method to normalize the lip location on images. The proposed method is based on a search algorithm of the lip position in which a lip location normalization is integrated in the model training [1] [4]. In order to study the proposed method, experiments based on HMMs (Hidden Markov Models) [2] on speaker-independent isolated word recognition were carried out using a lip centering approach as a preprocessing and a good recognition performance was achieved and the error rate was considerably reduced. Continuous speech recognition using lip location normalization also was improved.

2 Location Normalization

2.1 Location Normalized Training

This paper is based on a normalization method for the lip location which was presented in detail in [4]. As in [4], good recognition results for isolated word recognition were obtained. It is shown that the proposed method is able to normalize the position of the lips.

The method basically is composed of following two steps :

- **Best Location Search**
  For the training data set $\{I^{(k-1)}\}$, find the best lip location data set $\{I^{(k)}\}$ in the sense that the likelihood of each word in the data set is the highest for the corresponding HMMs$^{(k-1)}$, $(k = 1, 2, .., n-1)$.
- **Model Update**
  Update the model HMMs$^{(k)}$ by the Baum-Welch re-estimation algorithm using all training data set $\{I^{(k)}\}$ having the best location.

2.2 Lip Gravitational Centering

This experiment was carried out using images with lip tracking and an algorithm to locate the center place of the lips to bring them to the center place of frames.

For the experiments, the first frame of original utterances was taken and a image containing detected edges was built by sobel filtering. From this image, an algorithm was carried out in order to find the gravitational center coordinates (x_k, y_l), such that, the summation of intensities to the right and to the left are almost equal and also the summation of intensities to the top side and to the bottom side are almost equal.

After this, the lip-tracking procedure is followed in order to get the complete set of frames in which the lips are almost located in the center place of frames.

3 Experiments

3.1 Experimental Conditions

For experiments, the M2VTS database [5] was used and images from the lips were extracted after converting color images into gray-intensities images. Each frame consists of 80 × 40 pixels. The word boundaries of the training data were found by a HMM based speech recognition system which was used to segment and label the sentences.

Subsampling of data was carried out in order to reduce the computational time, that is, original frames were divided into blocks, each block having the average intensity of the pixels inside that block. For this study, subsampling data with blocks of 5 × 5 pixels, that is, two-dimensional feature vectors of 128 parameters were used. In order to solve the large variability of intensities, the average intensity over an utterance was subtracted from all pixels on it.

The lip location normalized training process was applied over the complete set of utterances. Experiments of recognition by using a continuous Hidden Markov Models (HMMs) were carried out. Each word model was represented by one HMM which is a left to right model with 8 states and two single Gaussian distributions of diagonal covariance. Each HMM model consisted of the static, delta and acceleration coefficients.

For the testing process, the leave-one-out method was used. It means that 37 leave-one-out testings were carried out, each one with 1440 training words. For the testing process, 40 test words per speaker, producing 1480 testing utterances in total.

Fig. 1 shows the experimental results. Words beginning in "w" means data attributes, such as, "T" data with lip tracking. "C" means that the lip gravitational centering algorithm has been applied over the data. "I" means intensity normalization. "k" means the iteration number of the location normalized training of HMMs$^{(k)}$ with "C", "T" and "I". The lip location normalized training process is also applied over the testing data set with intensity normalization. In case of iteration "0", the lip location normalization was applied over only the testing data.

The application of the lip gravitational centering algorithm is very effective, the error recognition rate was decreased by 4% for data with lip tracking and 4.3% for data with lip tracking and intensity normalization. In the experiments, a recognition rate of 76.9% and a reduction error rate of 18.5% were obtained at iteration "1". Although a recognition accuracy of 76.6% was obtained at iteration "3" in [4], this method gives a better result at iteration "1".

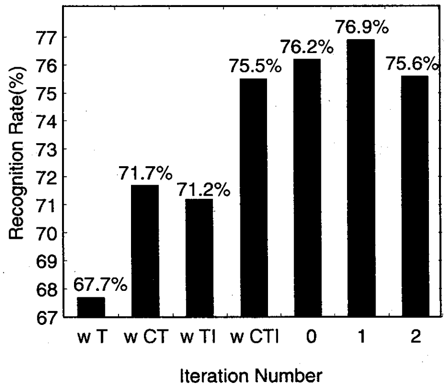

Figure 1: Effect of lip location normalization for images with subsampling data; blocks of 5 × 5 with lip tracking.

4 Continuous Speech Recognition

For continuous speech recognition two experiments were carried out. First, recognition experiments using H-MMs obtained from the natural movement and second, experiments using HMMs obtained by the lip tracking and lip location normalized training. In both cases subsampled data with 5 × 5 pixels was used. HMMs were used with 8 states and 2 single Gaussian distributions with diagonal covariance. The silence was modeled with 1 state and 2 Gaussian distribution mixtures and complete strings of images from zero to nine were built using lip images after applying the lip location normalization method. The normalized images were piled to make the string. Spaces between utterances were re-built after each utterance using same coordinates of the last image of each utterance.

Table 1: Continuous speech recognition results.

HMMs	w N	w I	"0"
Original Data	46.2%	52.6%	53.1%
	w T	w TI	"0"
Data with Location Normalization	56.8%	65.0%	66.2%

For data with original movement, a recognition accuracy of 52.6% was obtained for "w I" in table 1 for original data. 65.0% was obtained for "w TI" in case of data with location normalization. The location normalization method was applied for iteration "0" and a recognition accuracy of 53.1% was obtained for original data and 66.2% in case of data with lip location normalization. The lip location normalization is shown to be effective also in continuous speech recognition comparing the results with the ones obtained in [3].

5 Conclusions

In this study, we describe some experiments of isolated word recognition with normalization of the lip location. Prior the experiment, lip gravitational centering and a lip tracking algorithms were applied on images and better results were obtained.

Additional experiments of continuous speech recognition were carried out and it has been shown that a application of lip-tracking and normalized lip position to the original movement is effective also in continuous speech recognition.

References

[1] J.McDonough, T.Anastasakos and J.Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization," *ICASSP97*, pp.1043–1046, 1997.

[2] L.Rabiner and B.Juang, *"Fundamentals of Speech Recognition,"* Prentice-Hall, 1993.

[3] J.Luettin, "Towards Speaker Independent Continuous Speechreading". *Eurospeech97*, pp. 1991–1994, 1997.

[4] O. Vanegas, K. Tokuda, T. Kitamura, "Location normalization of HMM-Based lip-reading: Experiments for the M2VTS database", *ICIP99*, 26AP3.10. 1999.

[5] http://www.tele.ucl.ac.be/M2VTS/m2vts_db.ps.Z.