位置正規化学習を用いたバイモーダル話者認識

Oscar Vanegas, 徳田恵一, 北村正
名古屋工業大学

Bimodal Speaker Recognition using Location Normalized Training

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

1 Introduction

Is well known that in adverse conditions, the use of lip images for automatic speech recognition in addition of the acoustic speech signal, improves the recognition results. During the speech, however, the mouth region suffers some deformations, changing intensities on the lips area and sometimes allowing visibility of the teeth and tongue. This change contains information specific to the speaker and to the way that person speaks. Therefore, we are interested only in the mouth region for speaker verification and identification.

In this study, we used the image-based method, which is widely used due to its simplicity in its model construction. However, this method has the disadvantage of being affected very much for the variability of intensities and for the large variation of the lip locations. Therefore, this paper uses a method to normalize the lip location on images. In this paper, experiments based on HMMs (Hidden Markov Models) [1] on speaker recognition were carried out. The M2VTS [2] database was used for experiments. As a result of applying the described method, better speaker recognition performance was achieved and the error rate was considerably reduced.

2 Lip Extraction

The mouth region, normally presents deformations and the grey-level changes due to speaker dependent information. Based on this fact, a spatio-temporal model for each speaker was built. For the experiments, images from the lips in motion were extracted from the M2VTS database. The images were analyzed and a central point was visually calculated trying to leave the lips in the most center place of the frames.

For experiments, original full color images were converted into gray level images. Images of 80×40 pixels of the lips were cutted from original images. Two sets of data were used, one is the "original data" and the other "data with lip tracking and with location normalized training" [3] at iteration 3, which we will call in this paper to abbreviate "located data". Both type of data have intensity normalization in which the average value of intensities is subtracted from all the pixels in an utterance". The word boundaries of the data were found by an HMM based speech recognition which was used to segment and label the sentences. In order to reduce the computational time, images with subsampling data were used. For experiments, subsampling data with blocks of 5×5 pixels used 128 parameters.

3 Lip Location Normalization

As is well known, if lip images with a large variability of lip location are used for training, HMMs with a large variance can be obtained. Therefore, we presented in detail our lip location normalization method in [3]. This method is based on a search algorithm of the lip position in which a lip location normalization is integrated in the model training [4].

The normalization method is based on the two steps explained below :

- **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.

4 Speaker Recognition

4.1 Experimental Conditions

For Experiments, the M2VTS database was used. In this database, 37 people utter words from zero to nine in their native language. Five shots were recorded and the most difficult recording to recognize is the number five because of special characteristics are included in this shot.

For training data, the first 4 shots of the M2VTS were used and the fifth for testing data. One HMM

is trained for each digit and speaker. For the testing data, per user, the log likelihood was calculated for each digit using the HMM digit model of the claimed *k-th* person and the summation of likelihoods was finally computed.

Same calculation was carried out using the HMM digit model of each speaker except the claimed person. So that, finally we obtained the likelihood for the claimed *k-th* person and the average likelihood for the remaining persons. The difference between of the two scores is finally calculated as a normalized likelihood. A threshold value θ is calculated in which the number of false acceptance and false rejection is the minimum.

4.2 Acoustic and Labial Subsystem

The acoustic parameters are the Mel Cepstrum coefficients with static, delta and acceleration coefficients (39 components). Left to right HMMs were used with 8 states and only one single Gaussian mixture with diagonal covariance matrix. For the labial subsystem, each model consists of 3 states with one Gaussian distribution of diagonal covariance and the subsampling data was used. Parameters consist of 384 components composed of static, delta and acceleration coefficients.

For bimodal results, which is considered to be a very new domain [5], a normalization of scores was done by using the sigmoid function for both cases, acoustic and visual data.

$$\frac{1}{1 + exp(\eta * (x - \theta))} \quad (1)$$

where η is equal to -0.005, x is the calculated log likelihood and θ is the threshold value. In order to combine acoustic and image data, sigmoid functions for both data were linearly combined by weighting factor α and $1 - \alpha$.

4.3 Results

Tables 1 and 2 show the speaker identification accuracies, the false acceptance and false rejection rate for acoustic, original and located data. As in table 1, the located data showed better recognition results. The identification rate was increased from 70.3% up to 91.9%. As in table 2, combining both, the speech data and lip image data, the same tendency was shown. An identification accuracy of 100% was obtained for both, original and located data. 2.1% and 0% of false acceptance and false rejection were obtained using located data while 3.3% and 2.7% where obtained for original data respectively. Data with lip tracking and lip location normalization improved the false acceptance and rejection rates.

Table 1: Identification accuracies, false acceptance and false rejection rates given by original data and located data.

ID	%	FA	%	FR	%
Acoustic data					
37	100%	55	4%	1	2.7%
Visual : original data					
26	70.3%	230	16.8%	7	18.9%
Visual : located data					
33	91.9%	85	6.2%	2	5.4%

Table 2: Bimodal results.

ID	%	FA	%	FR	%	α
Visual : original data						
37	100%	46	3.3%	1	2.7%	0.9
Visual : located data						
37	100%	30	2.1%	0	0%	0.7

5 Conclusions

In this study, experiments of speaker recognition using acoustic and images from the lips, were carried out using the M2VTS database. We described a method to normalize the lip location on images from the lips in motion. Experiments were carried out in order to compare speaker recognition for original data and data with lip tracking and lip location normalization. It has been shown that the described location normalization is very effective for speaker recognition.

Acknowledgment

This work was partially supported by the Secom Science and Technology Foundation.

References

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

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

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

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

[5] Pierre Jourlin, Juergen Luettin, Dominique Genoud, Hubert Wassner, "Integrating Acoustic and Labial Information for Speaker Identification and Verification" *Proc. Eurospeech97*, pp.1603–1606, 1997.