

PARALLEL PHONE RECOGNIZER BASED MLLR SPEAKER RECOGNITION

Wang Eryu, Guo Wu, Dai Lirong

iFlytek Speech Lab, University of Science and Technology of China, Hefei
eryuwang@mail.ustc.edu.cn, {guowu,lr dai}@ustc.edu.cn

ABSTRACT

The method that uses maximum-likelihood linear regression (MLLR) adaptation transformation as features for support vector machine (SVM) has been adopted in recent NIST Speaker Recognition Evaluation (SRE). It is attractive because it makes use of high-level information about the speakers, and it can complement the standard GMM-UBM system. The performance of the system will be affected by the phone recognizer, especially in multi-lingual contexts. In this paper, we use a multi language phone recognizer based MLLR-SVM system, which can deal with the language phone recognizer problem. This system is defined as parallel phone recognizer-MLLR (PPR-MLLR). It has simpler framework than existing MLLR methods and can achieve better performance. In the NIST SRE 06 1conv4w-1conv4w task, the system can achieve an EER of 5.44%. Furthermore, we can achieve an EER of 4.20% which is almost a 20% system performance improvement when combined with the cepstral GMM-UBM system.

Index Terms—Speaker Recognition, Maximum-likelihood linear regression (MLLR), Support Vector Machine (SVM), Nuisance Attribute Projection (NAP)

1. INTRODUCTION

In text-independent speaker recognition, GMM-UBM is the dominant system, but the MLLR-SVM system [1,2,3] can provide complementary information to improve recognition rates. Compared with the traditional cepstral GMM-UBM system, the MLLR-SVM has solved two problems that [3] exist in cepstral systems. Firstly, it uses longer-term and higher level information in the speech by the introduction of phone, syllable or word models. Through this method, the speaker's speech style and other information can be reflected. Secondly, cepstral acoustic features depend on many factors, such as channel type, handset type, the spoken transcripts and language type. Though we can compensate channel difference with NAP [4, 5] for SVM system or factor analysis in GMM system, and diminish handset diversity with score normalization (ZNorm, Tnorm or ATNorm). But the transcripts and language factors cannot be avoided. MLLR-SVM system, which uses high level

information as its features, can work out the problem to some extent.

In speaker verification, a speaker may say words in any language while the MLLR-SVM depends on a given language recognizer. So it can achieve a better result if the test trials and training trials are both in the same language. It may face trouble when several languages are used. For example, the number of English trials is only about a half in NIST SRE 2006 core test [6], and there are also other language trials in the evaluation. Methods that ignore multi-language information do not perform well. To solve this problem, some researchers have introduced different recognizers in English trials and non-English trials [3]. For English trials, English tri-phone recognizer is used and for non-English trials, English mono-phone recognizer is used. But this is still cannot solve the problem entirely, due to the differences in different language phone models. One phone recognizer of one language cannot work well for all other languages. This is proved by the better performance in only English trials test than all trials.

In this paper, we propose a new MLLR-SVM system method to try to solve the problem of the language difference in multi-lingual speaker recognition. Instead of a tri-phone or four-phone acoustic model, we only use a mono-phone acoustic model which can not only greatly reduce time and cost in the decoding process, but also can be suitable for multi-language speaker verification. Here, we use several mono-phone acoustic recognizers in parallel. Like the framework of PPRSVM [7] system in language recognition, each language phone recognizer can do better in same language trials. When the results of these phone recognizers are combined together, some errors in one language phone recognizer can be revised in other phone recognizers. Through this method, we can make use of simple mono-phone recognizers to achieve a similar performance to other SVM-based systems and tri-phone model based MLLR-SVM systems [8]. To improve the performance of the system, we can combine parallel phone recognizer based MLLR-SVM systems with other systems. In experiments, we achieve a better result.

Section 2 firstly gives a short review of MLLR-SVM. Then we will introduce the flow of the PPR MLLR-SVM system in Section 3. Several experiments and contrastive experiments are presented in Section 4. And finally, conclusions and future work are given in Section 5.

2. MLLR-SVM SPEAKER VERIFICATION SYSTEM

MLLR-SVM depicts the characteristics of the speaker by the difference between the target speaker and UBM model. This difference can be figured out by the MLLR transform matrix in speaker adaptation. This transform can reflect the relationship of the speaker and UBM model in feature space location.

In maximum likelihood linear regression, an affine transform and a bias are applied in order to project the speaker-independent model to speaker-dependent model. This projection is usually used at Gaussian levels. The affine transform can be expressed in the pair of $[A, b]$ (A is a full affine matrix and b is a bias vector). In MLLR transforms, mean (μ) and variance (C) should both be updated. There are two types of MLLR transforms, one is constrained MLLR (CMLLR) [9], the other is unconstrained MLLR. CMLLR should use the same affine transform to update both mean and variance as follows:

$$\bar{\mu} = A\mu - b \quad \bar{C} = ACA^T \quad [1]$$

While unconstrained MLLR can use different affine transforms to update mean and variance respectively. Mean can also be updated independently.

Once the MLLR transform matrixes have been obtained, they can be sent to SVM as features. Other compensation methods such as NAP can also be applied to the MLLR-SVM system just like other SVM-based speaker verification systems.

3. PPR MLLR-SVM SYSTEMS

3.1. Parallel Phone Recognizer

There are six kinds of language phone recognizers in our PPR MLLR-SVM system. They work in parallel as shown in Fig.1.

Here, we introduce six language phone recognizers to solve the multi language problems. These phone recognizers are simple in framework. Unlike traditional MLLR-SVM systems, only mono-phone acoustic models are used here. Though the mono-phone acoustic model depends on specific language, it has less lexical and syntax information contained in the language, so it is fit for multi-lingual contexts. In our parallel phone recognizers based MLLR-SVM system, English, German, Hindi, Spanish, Japanese and Mandarin phone recognizers are used. They are trained by the OGI multi-language database. We label the data with worldbet[10]. IPA (International Phonetic Alphabet) is targeted at European languages and leaves out many of the sounds of the other languages which occur in the NIST SRE. Also, IPA does not label for the unusual sounds like clicks and noise. So worldbet labeling is more appropriate in our PPR MLLR-SVM.

The acoustic models are trained in 39 MFCCs acoustic features. And we can attain the transcripts of the speech data

from our phone recognizer. In the decoding process, one best result is adopted and a phone-loop net is used as a reference. Then we can use these phone level transcripts and mono-phone acoustic models to obtain MLLR and CMLLR transform matrices.

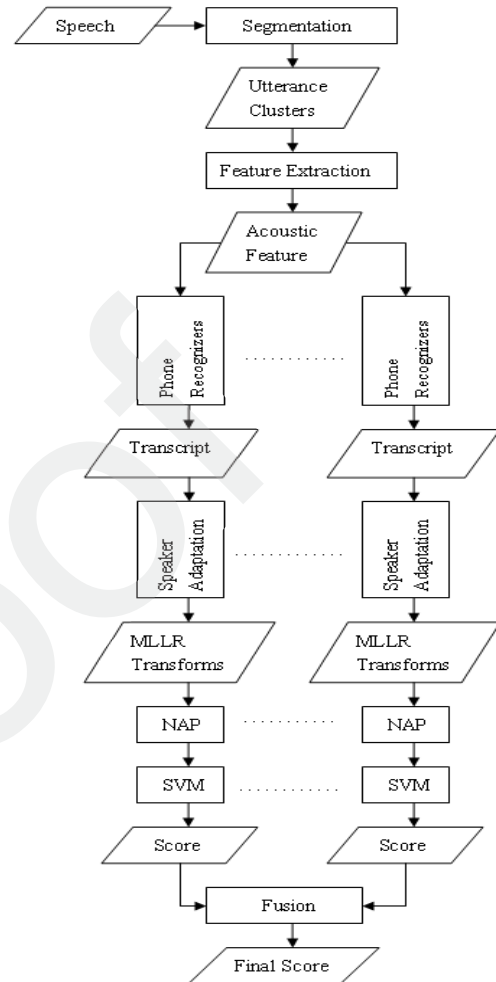


Fig.1 Flow chart of Multilanguage Phone Recognizer MLLR-SVM system

3.2. Speaker Adaptation Training

The MLLR transformation matrix can be shared among similar Gaussians. If we know the KL distance relationship among the mono-phone acoustic models, we can obtain appropriate phone sets. Each phone set shares the same MLLR transform matrix which can reflect the common characteristics that exist in the phone set. In our system, we use one CMLLR transform and two MLLR transforms (only mean of Gaussians updated). There are two passes in MLLR transforms extraction. The first pass is based on a phone-loop model, and we use 2 transforms for non-speech and speech respectively. The second pass uses a more detailed MLLR scheme. We use 3 transforms for non-speech, vowels and consonants. In the first pass, CMLLR transforms are

used while MLLR transforms are used in the second pass. The transforms for non-speech have no sense for speaker recognition. So we use other three transforms in our system.

3.3. MLLR Feature Extraction

Each speech file can be processed through the phone recognizer. Through the MLLR adaptation, there are 2 CMLLR and 3 MLLR transforms gained for the given speech. Non-speech sections are not helpful for speaker verification. We drop 1 CMLLR and 1 MLLR transforms of non-speech. Then, the coefficients of the other 3 transforms are concatenated. So a $(39*39+39)*3=4680$ dimensional supervector is attained. In order to get good classifiers, a feature-based normalization should be added to the supervector. Here, rank normalization is used to bring each dimensional feature to $[0, 1]$ by its rank between background distribution.

3.4. Nuisance Attribute Projection (NAP) and Score Normalization

NAP is a method to compensate the negative effect caused by channel difference. A development set is chosen to depict the channel characteristic. If the main eigenvectors of channel space are subtracted out of the original supervector, a less channel-dependent supervector is obtained.

When scores are obtained through the SVM, T_{norm} is applied to these raw scores. Score normalization is a score level compensation method which is always used to scale the raw scores.

3.5. Score Fusion

There are 6 sub-systems in our PPR MLLR-SVM system. Each score should be more credible for the same language trails. If the six scores are combined in a proper way [8], each phone recognizer should be more effective than when it is used alone. By doing this, multi language problems can be solved to some extent. But these weights need to be set by the development set to improve system performance. Here, we try to fuse the scores in a simple way. The average score is obtained instead of a weighted sum. The average score is not only representative of the six scores, but also independent of the development set. In our system, this method has been proved effective.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We used our PPR MLLR-SVM system in the NIST SRE 06 core task. In this task, there are 810 target speakers and 51448 trials. Training and testing speech data had a duration of about 5 minutes. This was cut to 2.5 minutes after silent frame were edited out.

We used 4555 conversation sides in NIST SRE 04 to train NAP transforms. These conversations were also used as a background training set. These data did not exist in the

NIST SRE 06; they also not the same as the speaker in NIST SRE 06.

645 conversation sides (371 for female and 274 for male) in SRE 05 were chose as the T_{norm} development set. We used gender-dependent normalization method.

All data was processed by our phone recognizers as described as above.

4.2. PPR MLLR-SVM System Results

As shown in Fig.1, we used 6 language phone recognizers in our system. For each phone recognizer, we evaluated its performance. The results of these phone recognizers are showed in Table 1:

Table 1 EER performance of 6 phone recognizers

	Baseline	T_{norm}
English	8.98%	8.69%
German	9.05%	8.91%
Mandarin	9.22%	8.92%
Spanish	9.86%	9.26%
Hindi	9.35%	9.10%
Japanese	9.42%	9.24%

From Table 1 we can see that each phone recognizer cannot achieve good performance by itself. Among these results, there English phone recognizer showed a little better performance than others. As we have described, for the core test which English trials take a large part of whole test, English recognizer is believed to work well.

Our PPR MLLR-SVM can greatly improve this performance by simple fuse these 6 phone recognizers. Table 2 shows this better result. Due to the difference language information supplied by multi-language phone recognizers, combination has showed a great improvement.

Table 2 EER performance between some SVM systems

System	Baseline	T_{norm}
PPR MLLR-SVM	5.70%	5.44%
GMM-SVM-MFCC	5.44%	4.89%
PPR MLLR-SVM+ GMM-SVM-MFCC	4.56%	4.20%

4.3. System combination

We compare our PPR MLLR-SVM system with one state-of-the-art cepstral SVM system. GMM-SVM system [11] has showed good performance in NIST SRE. This cepstral system uses 39 dimension MFCCs and a 256 mixture GMM-UBM model. Gender-dependent UBM models are adopted here. Then speaker model can be obtained by MAP adaptation. Finally, mean vectors of Gaussian Mixture in each speaker model are connected to a super vector. These vectors can be classified by a SVM model. NAP and T_{norm} are applied to this cepstral system.

In Table 2, the results show that baseline of PPR MLLR-SVM is similar to GMM-SVM system. But result with T_{norm} is less effective. Because of the similar scale of these two systems and the different features used, system combination should be impressive. When scores of the

cepstral systems were combined with the scores of the PPR MLLR-SVM system using average weight, the final result was proved to be best of all the systems. Combination between PPR MLLR system and GMM-SVM system can make EER of 4.20%. The obvious improvement was obtained because of the complementary information between two systems. The detect error tradeoff (DET) figure will show the performance among these systems.

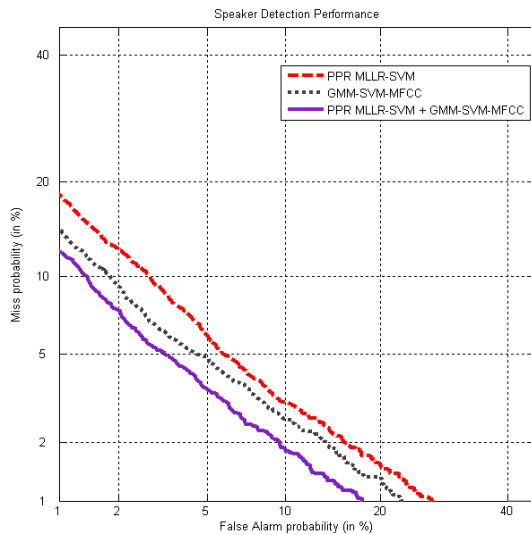


Fig.2 Detection Error Tradeoff Curve

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a new framework for MLLR-SVM systems. Several phone recognizers are introduced in our PPR MLLR-SVM system. In experiments, it works well in all language trials. At the same time, the performance of English trials is also improved. Although, each separate phone recognizer can not obtain an exciting result, PPR MLLR-SVM can greatly improve the final result when several phone recognizers work together. That means there is a close relationship between different phone recognizers. One phone recognizer could be effective for some trials, and another phone recognizer would be more appropriate for other trials. When their results are combined, a better result should be obtained. As we have described, this method works effectively in multilingual situations.

Compared with other N-gram MLLR-SVM systems, PPR MLLR-SVM has a much simpler framework. The framework of PPR makes use of mutual information among several phone recognizers and has a less time-consuming and simpler decoding structure. It also shows great improvement when combined with cepstral GMM-SVM systems.

Although the current system performance has reached the level of the normal cepstral system, there is still work to be done. Each phone recognizer is not so effective when they work alone. If performance of each phone recognizer can be improved, it is believed that the PPR system would

show a better result. We will try to introduce discriminative training and language model in phone recognizer training to increase its performance.

5. REFERENCES

- [1] A. Stolcke, L. Ferrer, S. Kajarekar, and A. Venkataraman, "MLLR transforms as features in speaker recognition", in *Proc. Interspeech*, Lisbon, pp. 2425-2428, Sep. 2005.
- [2] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-Transform-based Speaker Recognition", in *Proc of IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.
- [3] A. Stolcke, S.S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker Recognition With Session Variability Normalization Based on MLLR Adaptation Transforms", *IEEE Transaction on Audio, Speech, and Language Processing*, VOL. 15, NO. 7, Sep. 2007
- [4] A. Solomonoff, W. Campbell, and I. BoardmanCampbell, "Advances in channel compensation for SVM speaker recognition", in *Proc. ICASSP*, Philadelphia, PA, vol. I, pp. 629-632, Mar. 2005
- [5] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and nap variability compensation", in *Proc. ICASSP*, Toulouse, France, pp. 97-100, 2006
- [6] The NIST Year 2006 Speaker Recognition Evaluation Plan, available: <http://www.nist.gov/speech/tests/spk/2006>
- [7] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification", in *Proc of IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.
- [8] N. Brummer, L. Burget, and et al, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, NO. 7, Sep. 2007
- [9] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech Language*, 12, 75-98, 1998
- [10] J.L. Hieronymus, "ASCII phonetic symbols for the World's languages: Worldbet", *AT&T Bell Labs, Technical Report*, New York, USA, 1994
- [11] M.H. Liu, B.Q. Dai, Y.L. Xie, Z.Q. Yao, "Improved GMM-UBM/SVM for speaker verification", in *Proc. ICASSP*, Toulouse, France, 2006