CONFIDENCE MEASURE BASED MODEL ADAPTATION FOR SPEAKER VERIFICATION

Erhan MENGUSOGLU
Faculté Polytechnique de Mons – Multitel Research Center
Avenue Copernic, 1, 7000 Mons
Belgium

Abstract

Confidence measures are expected give a measure of reliability on the result of a speech/speaker recognition system. Most commonly used confidence measures are based on posterior word or phoneme probabilities which can be obtained from the output of the recognizer. In this paper we introduced a linear interpretation of posterior probability based confidence measure by using inverse Fisher transformation.
Speaker adaptation consists in updating model parameters of a speaker independent model to have a better representation of the current speaker. Confidence measures give a more reliable selection criteria to select the utterances which best represents the speaker. A linear interpretation of confidence measure is very important to select the most representative data for adaptation.

Key Words

Speaker Verification, Confidence Measure, Speaker Adaptation, Fisher Transformation.

1. Introduction

Improving the accuracy of a speech/speaker recognition system is a major research area in the current speech technology researches. Current speech/speaker recognition systems are not accurate enough when they are used in conditions different than the training conditions of the systems. The main research directions to improve the accuracy of systems are:

- Noise robust recognition
- Improved feature extraction
- Confidence measures
- Speaker model adaptation
- Improved language modeling

This paper is interested in the use of confidence measures in speaker model adaptation for the speaker verification task. Confidence measures are used on various applications in speech recognition field. In [1] and [2] confidence measures are used to spot erroneous words and to give a measure for word correctness in large vocabulary continuous speech recognition. In [3] confidence measures are used in a dialog system to improve the efficiency of dialog. In [4] application of confidence measure in language identification task is explained.
Confidence measures are generally based on posterior probabilities of the recognition system. Use of some prior information [5] has shown that improve the efficiency of confidence measures. Some confidence measures are based on language model probabilities [6]. In this case a word graph is constructed from language model probabilities and the word sequence obtained by the recognizer is scored by this graph.

2. Speaker Model Adaptation for Speaker Verification

Speaker verification is one of the useful biometric techniques which improve the security for access control. Here, we give descriptions and some applications of main techniques for adapting speaker models to changes in verification environment. These changes includes intra-speaker variabilities (speaking rate, changes related on sickness, …) and extra-speaker variabilities (noise, echo, …). When there is a mismatch between training conditions of speaker models, the verification system is less accurate.
The speaker verification technique used in this paper is Gaussian Mixture Model (GMM) [7] based and is not dependent on text content of the speech.
The structure of speaker verification system used is given on Figure 1 and 2. There are two types of model; speaker model and world model. Speaker model is trained on few sentences uttered by a speaker and world model is trained by a relatively large data set obtained from different speakers.
In the experiments, one state and two state GMMs are used. There are two types of two state GMMs:

1. One state for silence and one state for speech
2. One state for unvoiced phonemes and one state for voiced phonemes.

Initial speech/silence labeling and voiced/unvoiced labeling is obtained by applying HMM/MLP speech recognition [8] on the world model training data to obtain phoneme probabilities for each frame. These probabilities are then used to label the data. After having obtained a two state world model, it is used to label speaker data, then the labeled speaker data is used to train speaker models.

In the verification system, accept/reject decision is taken by comparing the likelihood score [7] of the utterance given the model of claimed speaker with the likelihood score computed for an impostor model. There are two impostor model included in the system. One for female and one for male speakers.

The main speaker model adaptation techniques are Maximum A posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR).

2.1. MAP

Model adaptation using MAP involves prior knowledge about the parameter distribution of the model to be adapted. This prior knowledge is used as a base for the new model which should better model the newly observed data. There is a weighing factor based on the availability of adaptation data and the degree of mismatch between newly observed data and data used to train the initial model. The use of prior information prevents over fitting the original model to the observed data. Since the amount of observed data is small, over fitting can cause degradation in accuracy of the system.

In MAP adaptation, the main goal is to maximize an a posteriori function which is based on likelihoods and prior probabilities. The model parameters are updated to achieve this goal: [9]

$$\lambda_{MAP} = \arg\max_{\lambda} f(\lambda|O) \qquad (2.1)$$

By applying Bayes' theorem, this formula takes the following form.

$$\lambda_{MAP} = \arg\max_{\lambda} \frac{L(O|\lambda)P_0(\lambda)}{P(O)} \qquad (2.2)$$

In this formula *O* is the observation vector. $L(O|\lambda)$ is the likelihood of the observed data given the present model. *P(O)*, the a priori probability of the observed data, is omitted because it does not depend on the model. $P_0(\lambda)$ is the prior probability density function of the model.

For simplicity, MAP adaptation can be used to adapt only the means of Gaussians in a GMM. The update formula [10] to obtain adapted means for some observation data is defined as follows,

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \qquad (2.3)$$

where t is the weighing factor for prior knowledge, N is the occupation likelihood of the adaptation data defined as,

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) \qquad (2.4)$$

R is the number of states, T is the number of observation vectors, μ_{jm} is the mean parameter of the model to be adapted and $\bar{\mu}_{jm}$ is the mean of the observation data defined as,

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \qquad (2.5)$$

As can be seen from the update formula, when the likelihood of observation data is higher the amount of adaptation is also higher. MAP adaptation perform better when more adaptation data is available.

2.2. MLLR

The use of MLLR for model adaptation consists in producing a set of regression based transforms from some adaptation data. These transforms are then used to tune the parameters of the GMM to be adapted. MLLR transformations are generally only applied to means of Gaussian which are the most important components of a GMM to be updated when it is adapted to new conditions. [11].

The use of MLLR for mean transformation of a Gaussian mixture model consists in computing a transformation matrix from observations and then using it to obtain adapted means.

For observations of dimension n,

$$\hat{\mu}_s = W_s \xi_s \qquad (2.6)$$

where W_s is an n x (n+1) transformation matrix and $\xi_s = [w, \mu_{s1}, \ldots, \mu_{sn}]^t$ is the extended mean vector so that w=1 indicates that there is an offset and w=0 means no offset.

W_s is computed by solving the following equation;

$$\sum_{t=1}^{T}\sum_{r=1}^{R} L_{sr}(t)\Sigma_{sr}^{-1} o(t)\xi_{sr}^{T} = \sum_{t=1}^{T}\sum_{r=1}^{R} L_{sr}(t)\Sigma_{sr}^{-1} W_s \xi_{sr}\xi_{sr}^{T} \qquad (2.7)$$

$L_{sr}(t)$ is the occupation likelihood which is obtained from forward backward process.
Implementation issues could be found in [9].

2.3. Use of Confidence Measure for Unsupervised Adaptation

Confidence measure is used to determine whether to use a certain utterance for adaptation or not. Before the adaptation procedure, each utterance is tested for a certain confidence threshold. Confidence measure is based on likelihood ratios. That means likelihood score from speaker model is divided by likelihood score from impostor model.

$$\Lambda(X) = \log p(X|\lambda_C) - \log p(X|\lambda_{\bar{C}}) \qquad (2.8)$$

$p(X|\lambda_C)$ is the likelihood that utterance *X* belongs to the claimed speaker and $p(X|\lambda_{\bar{C}})$ is the likelihood that utterance does not belongs to the claimed speaker. λ_C is the speaker model and $\lambda_{\bar{C}}$ is the world (background) model.
$\Lambda(X)$ is then compared with claimed speaker Gaussian mean and impostor Gaussian mean. Those Gaussians are computed using the claimed speaker data and impostor data. If the value is between two means then the confidence measure is computed by; first, normalizing the likelihood value by the two Gaussian(2.10 and 2.11), then, the normalized values are transformed to correlation domain by using inverse Fisher transformation [12]. This transformation is generally used for determining a confidence interval for the correlation value between two data set.

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \qquad (2.9)$$

In this formula, *z* is the transformed value of correlation value *r*. *z* has a Gaussian (normal) distribution. Correlation value can give the importance of relation between two data sets.
The values obtained after transformation are the measures of relationship between likelihood ratio and claimed speaker and impostor.

$$z_{speaker} = \frac{\Lambda(X) - \mu_{speaker}}{\sigma_{speaker}} \qquad (2.10)$$

$$z_{imposter} = \frac{\Lambda(X) - \mu_{imposter}}{\sigma_{imposter}} \qquad (2.11)$$

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad (2.12)$$

$$cm = r_{speaker} - r_{imposter} \qquad (2.13)$$

cm is the final confidence score that can be used directly for speaker model adaptation. If the confidence score is negative, that means the utterance comes from an impostor. If the score is positive, then utterance seems to be pronounced by the true speaker. In this case confidence score determines the level of confidence. For adaptation it is better to use only the utterances with high confidence scores.

3. Experimental Setup and Results

Some results obtained with the techniques explained in previous section are given. The use of confidence measures for unsupervised adaptation and some results are also included.

Three different types of experiment are realized to test the effect of adaptation on the accuracy improvement of speaker verification system.
1. Use of 1 state GMM for each speaker and the world model.
2. Use of 2 state GMM, one for speech and one for silence
3. Use of 2 state GMM, one for voiced phonemes and one for unvoiced phonemes.

The POLYCOST [13] speaker verification database is used for experiments. For each speaker, there are three set of data: training, testing, and adaptation data sets. A white Gaussian noise added (SNR=15) version of each data set is also used. A large amount of data is used for "world model" training and two impostor data sets (one for females and one for males) with their noisy version are also selected from the database.
Since the results for female speakers and male speakers are different, test results are listed separately for females and males.
The training and testing procedure for 2 state (speech and silence) GMM is shown in figure 1 and 2.

Results obtained from tests are listed below. There are 6 test groups. For every test, there is an explanation followed by a table. The tables are identical in form. There are four columns in the tables, type of test, sex of speakers, fault rejection error rate and fault acceptance error rate.

1. Clean data, use of 1 state GMM (res), 2 state GMM for silence-speech (res-sil) and two state GMM for unvoiced-voiced(res-uv). Best results obtained with voiced-unvoiced modeling for female speakers and silence modeling for male speakers.

res:	female	3.46%	2.95%
res:	male	4.69%	4.12%
res-sil:	female	3.21%	2.70%
res-sil:	male	3.70%	3.04%
res-uv:	female	3.09%	3.02%
res-uv:	male	4.07%	3.54%

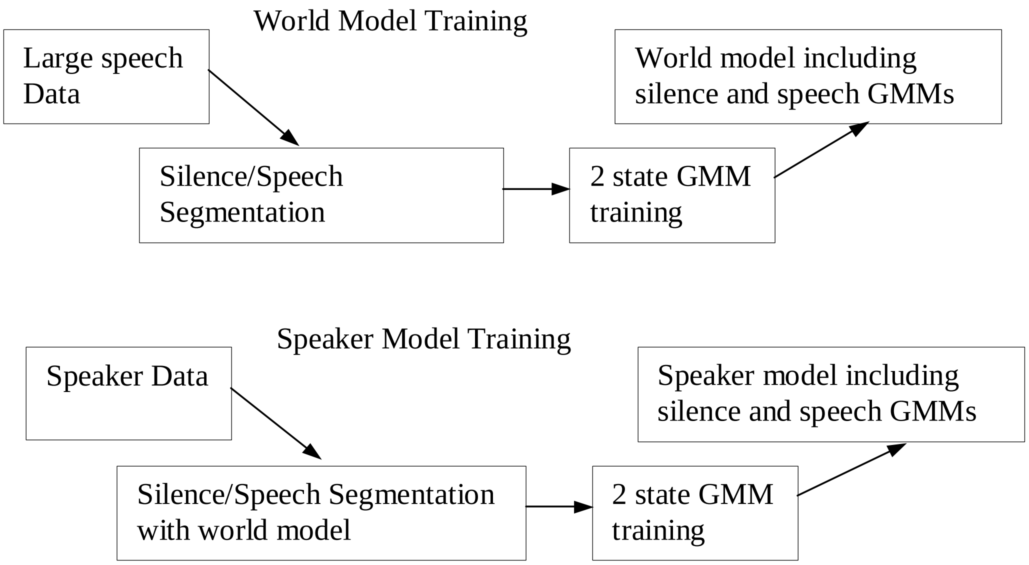
World Model Training
Large speech Data
Silence/Speech Segmentation
2 state GMM training
World model including silence and speech GMMs
Speaker Model Training
Speaker Data
Silence/Speech Segmentation with world model
2 state GMM training
Speaker model including silence and speech GMMs

Figure 1. Enrollment procedure for silence-speech modeling for Speaker Verification.

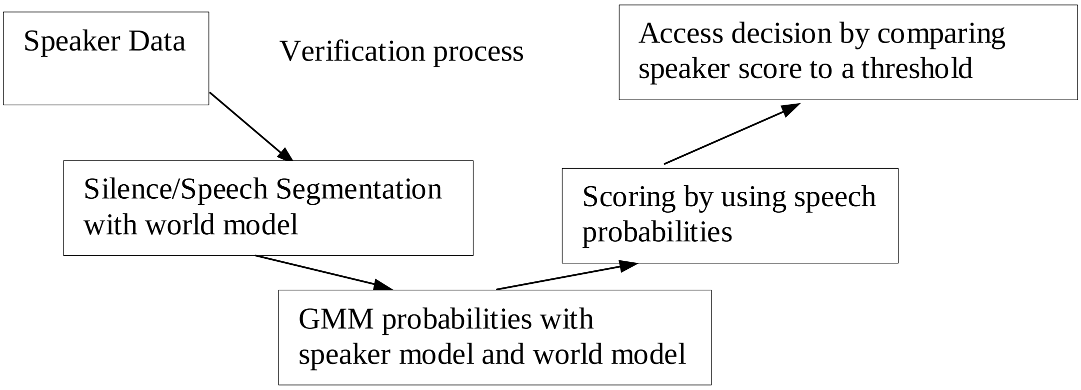
Speaker Data
Verification process
Access decision by comparing speaker score to a threshold
Silence/Speech Segmentation with world model
Scoring by using speech probabilities
GMM probabilities with speaker model and world model

Figure 2. Verification procedure for silence-speech modeling for Speaker Verification.

2. MLLR adaptation (resaall, res-silaall nad res-uvaall) and MAP adaptation (resaall-map) with clean adaptation data. Use of adaptation do not improve the results when there is no mismatch between training and test data.

resaall:	female	6.91%	5.44%
resaall:	male	5.06%	3.95%
resaall-map:	female	10.74%	8.34%
resaall-map:	male	10.25%	8.22%
res-silaall:	female	8.15%	6.45%
res-silaall:	male	4.44%	3.43%
res-uvaall:	female	7.04%	6.08%
res-uvaall:	male	4.44%	3.68%

3. Results with GMMs trained on clean data and tested on noisy(SNR=15) data. In noisy conditions, silence labeling is better for male speakers and 1-state modeling is better for female speakers.

res-n15:	Female	13.33%	9.82%
res-n15:	Male	9.14%	9.50%
res-sil-n15:	Female	16.91%	1.32%
res-sil-n15:	Male	8.64%	8.50%
res-uv-n15:	Female	14.44%	1.15%
res-uv-n15:	Male	10.00%	9.84%

4. Adaptation of GMMs trained on clean data to noisy data. MAP adaptation is only applied on single GMM case. MLLR adaptation give better results. Adapted models for male speakers give better results.

res-n15all:	female	13.58%	10.16%
res-n15all:	male	5.19%	3.73%
res-n15all-map:	female	14.57%	8.52%
res-n15all-map:	male	10.49%	8.59%
res-sil-n15all:	female	13.33%	12.79%
res-sil-n15all:	male	5.93%	4.39%
res-uv-n15all:	female	12.47%	10.51%
res-uv-n15all:	male	6.91%	4.54%

5. Use of confidence measure(confidence threshold=0.5) to select the adaptation utterance. (Only MLLR adaptation is used since MAP needs more data). Confidence measure driven adaptation improve the efficiency of female speaker models.

res-n15a:	female	12.10%	8.11%
res-n15a:	male	**5.56%**	**4.53%**
res-sil-n15a:	female	12.96%	11.44%
res-sil-n15a:	male	6.91%	5.84%
res-uv-n15a:	female	11.73%	9.23%
res-uv-n15a:	male	7.78%	6.07%

6. Results obtained for GMMs trained on noisy data. The results on the following table shows that confidence measure driven adaptation gives approximately same results as the GMMs trained on noisy data. In some cases the results are even better. For example in MLLR adapted single state GMM (bold in table 5 and 6).

res-n15-n15:	female	11.85%	7.04%
res-n15-n15:	male	**7.16%**	**4.20%**
res-sil-n15-n15:	female	13.82%	10.71%
res-sil-n15-n15:	male	9.51%	5.93%
res-uv-n15-n15:	female	13.95%	8.99%
res-uv-n15-n15:	male	8.77%	5.76%

4. Conclusion

The lowest error rates for clean data are obtained by using 2 state GMMs for speech and silence. Confidence measure driven MLLR adaptation perform better with 2 state GMMs in either speech-silence or voiced-unvoiced cases. MAP adaptation does not perform better than MLLR adaptation. Combined MAP+MLLR adaptation will be investigated. Use of the world model as a base for adaptation will be also tested. Another interesting test may be the use of varying prior knowledge weighing as a function of amount of data available. This could be done by decreasing the weight factor in MAP adaptation, t in (1.3), when there are more utterances to be used for adaptation. These utterances could be selected by confidence measure, which will prevent the over fitting of the model after adaptation.

5. Acknowledgement

This work is supported by Walloon Region through the “Initiative” ARTHUR project.

References

[1] F. Wessel, R. Schlüter, K. Macherey, H. Hey, Confidence Measures for Large Vocabulary Continuous Speech Recognition, *IEEE Transactions on Speech and Audio Processing, 9(3),* 2001, 288-298.

[2] S. Cox, High Level Approaches to Confidence Estimation in Speech Recognition, *IEEE Transactions on Speech and Audio Processing, 10(7),* 2002, 460-471**.**

[3] G. Skantze, The use of Speech Recognition Confidence Scores in Dialogue Systems, *Goteborg University, Graduate School of Language Technology, Speech Technology 1, course term paper,* 2003.

[4] F. Metze, T. Kemp, T. Schaaf, T. Schultz, H. Soltau, Confidence Measure Based Language Identification, *ICASSP 2000, Istanbul, Turkey,* 2000.

[5] E. Mengusoglu, C. Ris, Use of Acoustic Prior Information for Confidence Measure in ASR Applications, *EuroSpeech 2001, Aalborg, Denmark,* 2001.

[6] K. Hacioglu, W. Ward, A Concept Graph Based Confidence Measure, *ICASSP 2002, Orlando-Florida, USA,* 2002.

[7] D. A. Reynolds, Automatic Speaker Recognition Using Gaussian Mixture Speaker Models, *The Lincoln Laboratory Journal, 8(2),* 1995, 173-192.

[8] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach,* Kluwer, 1994.

[9] P. Nguyen, Fast Speaker Adaptation, *Technical Report, Eurecom,* 1998

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book for HTK Version 3.1* (Cambridge University Engineering Department, December 2001).

[11] Jonathan E. Hamaker, MLLR: A Speaker Adaptation Technique for LVCSR, *Lecture for a course at ISIP - Institute for Signal and Information Processing,* Department of Electrical and Computer Engineering, Mississippi State University, 1999.

[12] R. A. Fisher, *Statistical Methods Experimental Design and Scientific Inference* (Oxford Scince Publications, 1890-1962).

[13] D. Petrovska, J. Hennebert, H. Melin and D. Genoud, Polycost: A Telephone-Speech Database for Speaker Recognition, Speech Communication, 31(2-3), 2000, 265-270.