CLASSIFICATION MARGIN FOR IMPROVED CLASS-BASED SPEECH RECOGNITION PERFORMANCE

Denis Jouvet & Nicolas Vinuesa

Speech Group, INRIA - LORIA, 615 rue du Jardin Botanique, 54602 Villers les Nancy, France

ABSTRACT

This paper investigates class-based speech recognition, and more precisely the impact of the selection of the training samples for each class on the final speech recognition performance. Increasing the number of recognition classes should lead to more specific models, and thus to better recognition performance, providing the trained model parameters are reliable. However, when the number of classes increases, the amount of training data for each class gets smaller, and may lead to unreliable parameters. The experiments described in the paper show that taking into account a classification margin tolerance helps associating more training data to each class, and improves the overall speech recognition performance.

Index Terms— Speech recognition, class models, classification margin, speech classification.

1. INTRODUCTION

It is well known that many variability sources affect the speech signal and impact on the speech recognition performance [1], and that best speech recognition performance is achieved when operational (test) conditions match with the training conditions. That is why speech transcription systems used for transcribing audio signals runs in several passes, the first one being a diarization pass. The speech signal is split into segments according to the detected speaker changes, and, the environment condition (studio quality vs. telephone quality) as well as the gender are estimated for each segment. Then, in the following speech decoding pass, environment and gender specific models are used. Subsequent passes may be applied to refine the decoding through the use of discriminative models and unsupervised adaptation processes.

Increasing the amount of Gaussian components in the mixture densities usually improves the acoustic modeling through a better handling of various variability influences, and consequently improves the speech recognition performance. However because of the various variability values that need to be handled by the model (for example multiple speakers), the acoustic space covered by the mixture densities is rather large, and this limits the selectivity of the densities, and hence the recognition performance. One way to tackle this phenomenon is to use a multiple modeling approach [1]. Instead of having a single acoustic model covering all the variability values, several models are developed, each model covering only a subset of the variability values. Then for the recognition process several schemes are possible. The variability value can be estimated in a deterministic way and the corresponding model used for decoding the utterance, or the decoding can be performed for each model and the one leading to the best likelihood score provides the answer. Other combinations of multiple decoding answers are also possible, such as the ROVER approach [2].

When the speaker is known, speaker dependent modeling is the most efficient approach. Adaptation techniques are useful to derive good speaker dependent models from a generic speaker independent model and some adaptation data collected from the target speaker. When only a limited amount of adaptation data is available, acoustic models can be adapted through eigenvoice-based techniques [3] or through interpolating cluster-based models [4] or reference speaker models [5].

Dynamic Bayesian network (DBN) [6] provides an efficient framework for making acoustic models dependent on some auxiliary variable that represents a variability source under consideration, as for example the pitch in [7], some hidden factors as in [8] or some inter-speaker variability [9].

In those approaches, class (i.e. condition, gender, speaker, ...) specific models need to be trained. Ideally we would like to have specific acoustic models for each possible class in order to achieve performance similar to that provided by a speaker and condition adapted model. However, when the number of classes increases, there is less and less data available for training the class-model parameters, the model gets unreliable, and speech recognition performance degrades.

Hence the main topic of this paper which is the introduction of a classification margin in order to increase the amount of data associated to each class for an improved training (or adaptation) of the class-model parameters. This is, in some way, similar to the handling of the boundary uncertainty that was investigated in [10]. The handling of the uncertainty led to increasing the amount of examples used for estimating speaking rate dependent pronunciation variant probabilities.

The organization of the paper is the following. Section 2 recalls the class-based speech recognition approach and presents the procedure used for creating automatically the training classes. Section 3 introduces the classification margin feature, and analyses its impact with respect to the size of the classes. Section 4 presents the speech recognition experiments and discusses the results. Finally a conclusion ends the paper.

2. CLASS-BASED SPEECH RECOGNITION

Traditional approaches for transcribing broadcast news data rely on environment and speaker-gender specific models (e.g. [11]). Thus

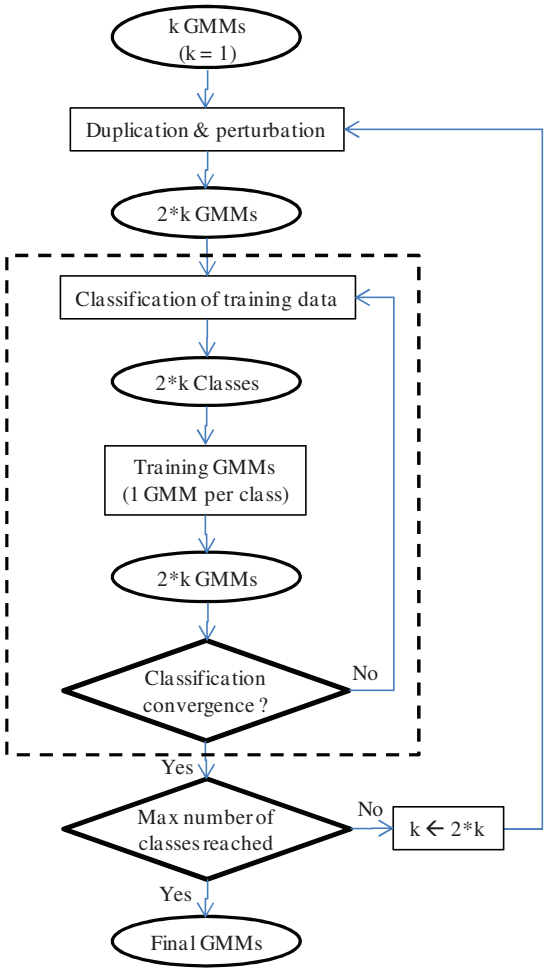

Fig. 1. - Iterative procedure for automatic training data classification

acoustic models are typically trained for studio quality data (8 kHz bandwidth) and telephone quality data (4 kHz bandwidth), and then adapted to the gender.

For transcribing the data, after a first diarization pass, the decoding of each audio segment uses the phoneme acoustic models corresponding to the estimated environment and speaker-gender.

2.1. Speech recognition and class-based models

This aim of class-based speech recognition is to extend the approach beyond the 2 traditional classes associated to the speaker gender.

So, let say we have K data classes ($C_1, C_2, ..., C_K$). A GMM Φ_k is associated to each class C_k, and used for classifying unknown audio segments X_i:

$$X_i \in C_k \Leftrightarrow P(X_i|\Phi_k) \geq P(X_i|\Phi_l) \qquad \forall l \qquad (1)$$

Once, an audio segment X_i is associated to a class C_k, the corresponding speech signal is decoded with the phoneme acoustic models Λ_k that have previously been trained (adapted) using data of the class C_k. The optimal sequence of words $\hat{W}$ is then given by:

$$\hat{W} = \arg\max_W P(X_i|W, \Lambda_k) P(W) \qquad (2)$$

2.2. Automatic classification of training data

In this paper, we have used an automatic procedure to classify the training data in a arbitrary number of classes, in fact, 2, 4, 8 or 16 classes. The procedure works in an iterative manner, as presented in Fig. 1. The acoustic analysis is the same as for speech recognition (MFCC features, plus first and second temporal derivatives).

The procedure starts with a single class corresponding to the whole training data, and first estimates the corresponding GMM (k=1), top of Fig. 1.

Then at each pass, the GMMs are duplicated (i.e. GMM Φ_k is copied as Φ_{k_1} and Φ_{k_2}) and the values of their mean parameters are modified by a small random value. Then,

1. The whole training set is classified using this new set of GMMs (each audio segment being associated to the class/GMM that maximizes the likelihood as in EQ (1)).
2. The data in each class is then used to train the corresponding GMM.
3. The above classification and training steps are repeated until a convergence criterion is reached (or maximum number of iterations).

The number of GMMs (one per class) is increased up to a predefined number of classes.

3. CLASSIFICATION MARGIN

In a traditional approach, each audio segment of the training set is associated to a single class, that is to the class corresponding to the GMM which provides the highest likelihood.

However, the characteristics of the data at the boundaries changes slowly when moving from one class to the next one. So, it seems reasonable to affect data which are at the boundary of adjacent classes to the two classes.

3.1. Classification margin

The boundary between two classes corresponds to the audio segments which have the same likelihood with respect to the GMMs of these two classes. Hence EQ (1) can be modified in order to associate audio segments which are close to the boundary to several classes:

$$X_i \in C_k \Leftrightarrow P(X_i|\Phi_k) \geq \max_l P(X_i|\Phi_l) - \delta \qquad (3)$$

where δ is the tolerance margin for classifying audio segments.

When δ is set to 0, EQ (3) and (1) lead to the same classification results. Increasing the margin δ increases the amount of audio segments which are associated to any given class.

3.2. Experimental analysis of some classes

Figures 2 and 3 display the number of audio segments of the training data which are associated to each class for different values of the classification margin δ. Logarithmic scales are used on the

vertical axes. On Fig. 2, dashed lines are used for the 2-class approach, and solid lines are used for the 4-class approach. Fig. 3 display the corresponding information for the 16-class approach.

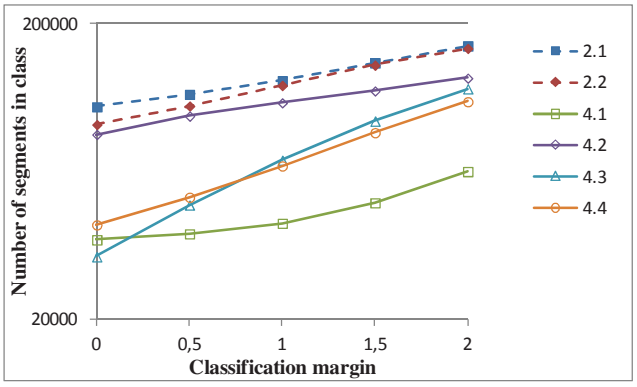

Fig. 2. - Number of audio segments for training the acoustic models of each class of the 2-class and 4-class models according to the classification margin.

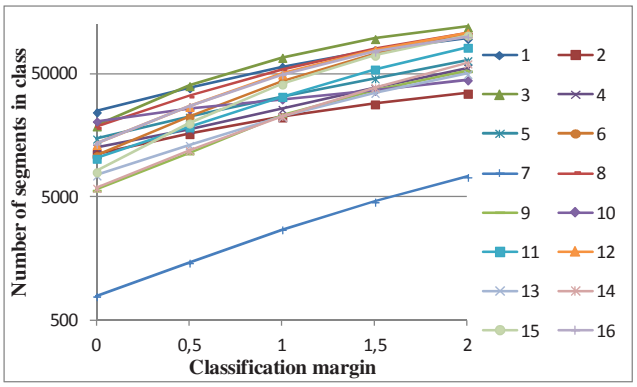

Fig. 3. - Number of audio segments for training the acoustic models of each class of the 16-class models according to the classification margin.

The figures show that the amount of data is not evenly distributed between the classes (although random perturbations of the GMM parameters are applied when increasing the number of classes). In the 4-class approach, one class has much more data than the three other ones, and in the 16-class approach, one class has one order of magnitude less data than the other classes.

All the curves show a smooth increase of the number of audio segments associated to each class when the classification margin value is increased from 0.0 up to 2.0.

4. SPEECH RECOGNITION EXPERIMENTS

The speech recognition evaluations have been conducted using French broadcast news data from the ESTER2 evaluation campaign [12].

4.1. Speech recognition framework

All the experiments have been conducted using the Sphinx speech recognition toolkit [13]. Moreover, as we focused here on the classification impact, we apply only a single pass speech decoding. This means that there is no discriminative processing (LDA, MPE, ...) nor speaker adapted acoustic models (MLLR, SAT, ...) used. Hence, after the diarization step which segments the audio data according to speakers, and classifies each audio segment with respect to the environment (studio quality data vs. telephone quality data), the class-based speech decoding is applied for each segment: the class corresponding to the segment is determined (highest GMM likelihood - i.e. each segment is associated to a single class), and then, the speech decoding is performed with the phonetic acoustic models corresponding to that class.

Each acoustic model has 4500 senones (shared densities) and 64 Gaussian components per mixture. Generic phonetic models (one for studio quality, and one for telephone quality) are first trained using all the available training data. Then, the context-dependent acoustic models are adapted to each class using the associated data, as defined by EQ (3). Hence the training data selected for training the class acoustic models takes into account the classification margin δ. When the margin δ gets larger, more audio segments are selected for adapting the class acoustic models.

The training is carried out on the ESTER2 training data (about 190 hours) and the recognition results are reported for a large subset of the ESTER2 development data, about 4h30 of audio signal corresponding to 36 800 running words.

In our experiments, the pronunciation lexicon used for speech recognition contains about 64 000 entries. A trigram language model is used.

4.2. Speech recognition performance evaluation

Word error rates measured on the ESTER2 development data are reported in Table 1. for different class configurations (2, 4, 8 and 16 classes), and taking into account different classification margin values for classifying the training set.

Table 1. Word error rates on ESTER2 development data with respect to the number of classes and the classification margin used in selecting the training data for each class.

Margin	No margin	0.5	1.0	1.5	2.0
2 classes	25.24%	**25.06**%	25.21%	25.60%	25.65%
4 classes	25.12%	25.07%	**25.04**%	25.20%	25.43%
8 classes	25.05%	24.95%	24.93%	**24.81**%	25.14%
16 classes	25.66%	24.76%	**24.67**%	24.77%	25.11%

When only the generic environment models (i.e. studio quality and telephone quality) are used, which amounts to having only 1 class, the word error rate achieved on this ESTER2 development data is 25.97%. When the traditional gender classification (male vs. female adapted models) is used, the word error rate goes down to 24.91%.

The results displayed in the column "no margin" correspond to a standard classification of the training data, i.e. each training audio segment is associated to the class of the GMM leading to the highest likelihood. With 2 classes, there is a significant improvement with respect to the 1-class approach, but the results

are not as good as those obtained with the gender specific models. Moreover, when the amount of classes gets too high, speech recognition performance degrades.

Each line of the table shows the impact of increasing the training set of each class through the classification margin δ. Each line exhibits a similar behavior. By introducing a classification margin when associating the training audio segments to each class, this increases the number of audio segments associated to each class, as discussed previously in section 3.2. Having introduced rather similar additional data (at least when the margin is not too large), this does not affect too much the relevance of the estimated parameters with respect to the acoustic modeling of the class. However, having enlarged the size of the class training set makes the resulting class model parameters more reliable. But using a too large margin (e.g. 2.0 in Table 1) degrades the speech recognition performance. The optimal values range between 0.5 and 1.5 depending on the number of classes used.

It is also interesting to note that by improving the training of the class acoustic model parameters, it is possible to successfully go beyond the traditional gender-based classification, and outperformed its speech recognition performance (24.67% for the 16-class approach vs. 24.91% for the gender-based approach).

In the reported experiments, only MAP adaptation was used to adapt the generic model parameters to each class data. Its combination with MLLR has not yet been investigated. Also, as the amount of data associated to each class is highly variable, it might be interesting to adjust the classification margin according to the class population in order to guarantee enough data for a reliable adaptation.

Finally, another important aspect to investigate further is the classification of the training data. Here a simple technique based on GMM modeling was used. More refined speaker similarity metrics have been investigated in the past (e.g. [14], with methods emphasizing on vowels, or methods using phonetic HMMs, ...) and could be used to cluster the training data and build the initial classes.

5. CONCLUSION

In this paper we have investigated the training of class-based acoustic models in a speech transcription context. A standard automatic classification procedure has been presented to build automatically an arbitrary amount of classes. Here, 2, 4, 8 and 16 classes were created. The approach iterates duplication and perturbation of GMM models, followed by several classification and training steps. When the number of classes increases, there are fewer and fewer training data in each class for adapting the acoustic class-model parameters, which may lead to unreliable parameters and thus degrade speech recognition performance.

Hence the main point investigated in the paper, which is the introduction of a classification margin in the selection of the training data associated to each class. With such a margin, the selected data comes from the class itself (audio segments with highest likelihood obtained for the GMM of the class), and also from similar data belonging to other classes, but close to the class boundary. This way, more data are selected for adapting the parameters of each class acoustic models. This helps obtaining more reliable acoustic models for each class. The experiments have shown that using a small classification margin improves the speech recognition performance. This way, it was possible to go beyond and to outperform the traditional gender-based classification.

Future work should investigate more refined classification techniques in order to achieve more homogeneous classes, and improve further the speech recognition performance.

6. REFERENCES

[1] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi & C. Wellekens, "Automatic speech recognition and variability: a review", *Speech Communication*, vol. 49, pp. 763-786, 2007.

[2] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", *Proc. ASRU'97*, Santa Barbara, CA, USA, pp. 347-354, 1997.

[3] R. Kuhn, P. Nguyen, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field & M. Contolini, "Eigenvoices for speaker adaptation", *Proc. ICSLP'98*, Sydney, Australia, pp. 1771-1774, 1998.

[4] M.J.F. Gales, "Cluster adaptive training for speech recognition", *Proc. ICSLP'98*, Sydney, Australia, pp. 1783-1786, 1998.

[5] T. Wenxuan, G. Gravier, F. Bimbot & F. Soufflet, "Rapid speaker adaptation by reference model interpolation", *Proc. INTERSPEECH'2007*, Antwerp, Belgium, pp. 258-251, 2007.

[6] G. Zweig, "*Speech recognition with Dynamic Bayesian Networks*", Ph. D. Dissertation, Univ. California, Berkeley, 1998.

[7] T. A. Stephenson, M. Magimai-Doss & H. Bourlard, "Speech recognition with auxiliary information", *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 189-203, 2004.

[8] F. Korkmazsky, M. Deviren, D. Fohr & I. Illina, "Hidden factor dynamic Bayesian networks for speech recognition", *Proc. ICSLP'2004*, Jeju Island, Korea, 2004.

[9] G. Cloarec & D. Jouvet, "Modeling inter-speaker variability in speech recognition", *Proc. ICASSP'2008*, Las-Vegas, USA, March 2008.

[10] D. Jouvet, D. Fohr & I. Illina, "About handling boundary uncertainty in a speaking rate dependent modeling approach", *Proc. INTERSPEECH'2011*, Florence, Italy, Aug 2011.

[11] P. Deléglise, Y. Estève, S. Meignier & T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?", *Proc. INTERSPEECH'2009,* Brighton, UK, Sept. 2009.

[12] Galliano, S., Gravier, G., and Chaubard, L., "The Ester 2 evaluation campaign for rich transcription of French broadcasts", *Proc. INTERSPEECH'2009,* Brighton, UK, pp. 2583-2586, Sept. 2009.

[13] (2011) Sphinx. [Online] Available: http://cmusphinx.sourceforge.net.

[14] S Krstulovic, F. Bimbot, O Boëffard, D. Charlet, D. Fohr & O. Mella, "Selecting representative speakers for a speech database on the basis of heterogeneous similarity criteria", *Speaker Classification II*, Christian Müller (réd), *Lecture Notes in Computer Science*, 4441, Springer Berlin, pp. 276-292.