MODELING INTER-SPEAKER VARIABILITY IN SPEECH RECOGNITION

Gwenael Cloarec & Denis Jouvet

France Telecom – Division R&D – TECH/SSTP
2, Avenue Pierre Marzin, 22307 Lannion, France

ABSTRACT

This paper details a method for taking into account variability influence in HMM-based speech recognition. The set of Gaussian components of the mixtures represents the entire acoustic space covered for all possible variability values. For each utterance to be recognized, the corresponding variability value is estimated and used to weight and/or constrain dynamically the acoustic space for each pdf. To do that, the weight coefficients of the Gaussian mixtures are set dependent on the variability value. As an example, the variability considered is the inter-speaker variability, and is handled through speaker classes. Taking into account for each utterance the four speaker classes that best match with the utterance signal leads to a significant word error rate reduction on a continuous speech recognition task, as compared to standard speaker-independent modeling.

Index Terms— Speech recognition, acoustic modeling, dynamic Bayesian network, inter-speaker variability, speaker class.

1. INTRODUCTION

It is well known that HMM-based speech recognition systems are very sensitive to sources of variability that affect the speech signal. This is why best performances are achieved when test (operational) environment matches with the training environment. Also, the more constrained the operating conditions are, the smaller the variability of the speech signal is, and consequently the better the recognition performance is. This is why speaker-dependent systems provide better recognition performance than speaker independent systems. However using a speaker-dependent system is not a tractable solution in voice interactive services where anybody can call the service.

For speaker independent systems, the acoustical models are usually adapted to the operational environment, using field speech data collected from actual interactions between the speakers and the vocal service. Thus, the field adapted model matches as closely as possible the operational condition, but large variability variations still needs to be handled; they are due to many factors [1] such as inter-speaker variability, varying noise level, etc.

Increasing the amount of Gaussian components in the mixtures usually improves the acoustic modeling, and consequently 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 pdfs is rather large, and limits the selectivity of the densities, and hence the recognition performance. One way to handle 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 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 score provides the answer. Other combination of multiple decoding answers is 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 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 represent the variability source under consideration, as for example the pitch in [7] or some hidden factors as in [8].

The approach proposed in this paper benefits from several of the above techniques for handling inter-speaker variability. First several classes of speakers are determined from the training set data, and acoustic models are trained for each class of speakers, hence providing multiple acoustic models. The acoustic models are merged at the acoustic level, this means that the Gaussian components of each pdf mixture are obtained by pooling the corresponding Gaussian components of the various speaker-class models. Further the

weight coefficients of the resulting pooled mixtures are set dependent on a speaker class variable using a DBN framework. Then before recognizing an utterance, the speaker classes that best match with the utterance are determined and used to adjust the weight coefficients of the mixture components.

The paper is organized as follows. Section 2 details the modeling of inter-speaker variability. Section 3 presents the experimental setup. Section 4 analyses the recognition results. Finally conclusions are drawn in section 5.

2. MODELING INTER-SPEAKER VARIABILITY

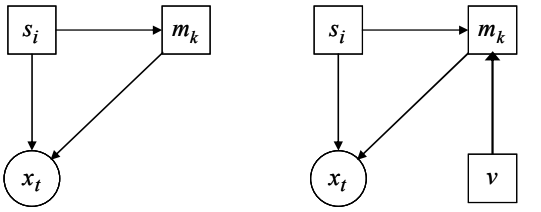
The Dynamic Bayesian Network formalism clearly exhibits the dependencies that are taken into account in the acoustic modeling. In Figure 1, the left part (a) represents the classical Gaussian mixture modeling. The mixture component m_k depends on the state s_i, and the probability of an observation vector x_t is given by:

$$P(x_t|s_i) = \sum_{k=1\cdots K} P(m_k|s_i)P(x_t|m_k,s_i) \qquad (1)$$

The right part (b) describes the proposed approach. The weight coefficient of the Gaussian component m_k is set dependent on a variability variable v. This leads to the following observation probability:

$$P(x_t|s_i) = \sum_{k=1\cdots K} P(m_k|s_i,v)P(x_t|m_k,s_i) \qquad (2)$$

In the following, the variability under consideration refers to the inter-speaker variability. However, the approach translates easily to any kind of variability.



(a) Classical Gaussian mixture (b) Variability dependent Gaussian mixture

Figure 1: *Introducing dependence on variability source.*

Inter-speaker variability is handled through classes of speakers which exhibits similar characteristic. A clustering technique, described in section 3, is used for obtaining those classes on the training data. Let assume that N speaker classes are determined on the training set. An acoustic model is then estimated for each speaker-class using the subset of training data corresponding to that class. As an adaptation technique is used for adapting a speaker independent model on each subset of training data associated to a class of speakers, it is reasonable to assume that each pdf of the various speaker-class adapted models represents the same part of sound as for the corresponding pdf of the speaker independent model. Hence, these various Gaussian components can be pooled to defined the entire acoustic space covered for the various variability values (here all the classes of speakers). This is represented in Figure 2, where $G_{s_i,v_j,k}$ corresponds to the Gaussian component $G_{s_i,k}$ adapted on the subset of training data corresponding to variability value $v = v_j$.

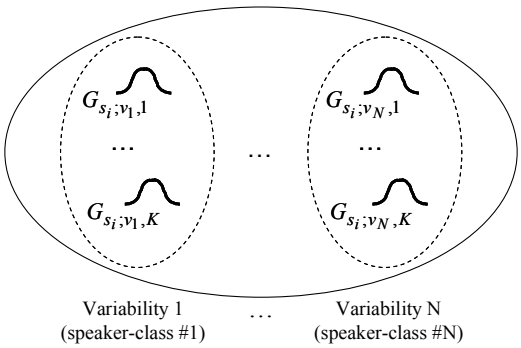


Figure 2: *Pooling Gaussian components associated to different variability sources.*

The resulting pdf relies on all these pooled Gaussian components, but weight them according to an estimation of the variability variable v for the current utterance X. This is done in the following way:

$$\begin{aligned} P(x_t|s_i) = {} & \lambda_{v_1}(X) \sum_{k=1\cdots K} c_{s_i,v_1,k} G_{s_i,v_1,k}(x_t) \\ & + \lambda_{v_2}(X) \sum_{k=1\cdots K} c_{s_i,v2,k} G_{s_i,v_2,k}(x_t) \\ & \cdots \\ & + \lambda_{v_N}(X) \sum_{k=1\cdots K} c_{s_i,v_N,k} G_{s_i,v_N,k}(x_t) \end{aligned} \qquad (3)$$

where $c_{s_i,v_j,k}$ are the weight coefficients of the adapted models

$$\sum_{k=1\cdots K} c_{s_i,v_j,k} = 1 \qquad \forall\ j = 1\cdots N \qquad (4)$$

and the weight coefficients λ_{v_j} are such that

$$\sum_{j=1\cdots N} \lambda_{v_j}(X) = 1 \qquad (5)$$

It should be noted that some weight coefficients can be set to zero, thus forbidding the use of the corresponding region of the acoustic space.

Although in Eq. (3) the weight coefficients λ_{v_j} are estimated using the entire utterance X, the approach translates to a local estimation of a variability criterion, either using the current frame, or the beginning of the utterance, or even a previous utterance of the same speaker.

It all depends on the variability under consideration, and the data necessary for its adequate estimation.

3. EXPERIMENTAL SETUP

Experiments were conducted on telephone speech data using an HMM-based speech recognition system. The training data came from the NEOLOGOS corpus available at ELDA [9]. The test data was collected during field experiments of the *Plan Restau* service, hence providing spontaneous continuous speech data.

The acoustic analysis was carried out with the front-end algorithm ETSI ES 202 212 [10]. This acoustic analysis was designed for providing a noise robust front-end for distributed speech recognition systems. Here, 10 MFCCs and the log Energy coefficients were used, together with their first and second order temporal derivatives. Baseline acoustic vectors are thus composed of 33 coefficients.

The NEOLOGOS corpus was produced within the French national project NEOLOGOS, as part of the Technolangue programme funded by the French Ministry of Research and New Technologies (MRNT). It was designed in order to represent inter-speaker variability and to be characteristic of the French population [11]. The main underlying objective was to select a limited amount of representative speakers that provided a good coverage of the French speakers, and to collect a large amount of data for each of these typical speakers. The speakers were selected on acoustic basis from preliminary speech data collected from a much larger set of speakers. The final speech corpus actually consists of 100,000 utterances of different nature (connected digits, telephone numbers, credit card numbers, spelled words, prompted names and phonetically rich sentences) corresponding to 200 selected speakers from the different regions of France.

The NEOLOGOS corpus was then used to define classes of speakers and for training the acoustic models associated to each class. The N (here $N=10$) classes of speakers were determined automatically through the following procedure:

- First context independent phone models relying on single Gaussian densities were estimated for each of the 200 speakers of the NEOLOGOS corpus.
- Then the Gaussian pdfs associated to the central states of oral vowels, nasal vowels, fricatives and nasals consonants were considered, and the Kullback-Leibler distance was computed between these central state densities corresponding to the different speakers.
- Finally the N speaker classes were built through hierarchical clustering [11].

As the clustering was performed on acoustic basis, the data in each class should come from speakers having similar characteristics. The size of the resulting classes varies from 6 speakers for the smallest one to 52 for the largest one.

Two acoustic models were trained on each class of speakers. One was devoted to speech recognition and the other to speaker-class recognition in order to estimate the variability value, and thus determine the weight coefficients of the Gaussian mixtures.

The speech recognition models were based on context-dependent modeling of the phonemes, with an a priori sharing of the Gaussian mixture densities between contexts having similar influence on the acoustic realization of the sounds [12]. First speaker independent acoustic models were estimated using the entire training set (i.e. 200 NEOLOGOS speakers). Mixtures with 8 Gaussian components were used. This was the baseline modeling.

Then using the training subset corresponding to each class of speakers, the generic baseline model was adapted for each class of speakers. This provided the $N=10$ speaker-class acoustic models. The Gaussian components of the corresponding pdfs were later pooled as described in section 2; and the weight coefficients were determined as described below.

Text-independent speaker recognition techniques were used to estimate for each utterance the "similarity" between the speaker that has uttered this utterance and each speaker-class,. A Gaussian mixture model (GMM) Λ_j was estimated on the subset of training data corresponding to each speaker-class j. Then the likelihood $P\left(X\middle|\Lambda_j\right)$ of the utterance X to be recognized was computed for each speaker-class model Λ_j.

Some recognition experiments were conducted using only the best matching speaker-class. This amounts to having $\lambda_{v_1}=1$ with v_1 corresponding to the best matching speakers-class index, and $\lambda_{v_j}=0$ for $v_j \neq v_1$. This is like decoding with only the acoustic models of the best matching class.

In the second set of experiments, the 4 best matching speaker classes were used with the following weight coefficients: $\lambda_{v_1}=0.4$, $\lambda_{v_2}=0.3$, $\lambda_{v_3}=0.2$, $\lambda_{v_4}=0.1$, and $\lambda_{v_j}=0$ for $j>4$, where v_1 corresponds to the best matching speaker-class index, v_2 to the second best matching class, and so on.

Also, for comparison purpose, a last recognition experiment was carried out using the acoustic models from all the 10 speaker-classes together. The acoustic models were merged at the pdf level by pooling the Gaussian components of the corresponding speaker-class model pdfs.

The test data came from the *Plan Restau* task. It is a continuous speech recognition task used with the spoken dialog system described in [13] for a tourism telephone service. This task is based on a vocabulary of 2200 words [14]. The test corpus consists of 1803 utterances collected

over the telephone from field experiments and corresponds to 7607 words. The language model is a bi-gram model.

4. RESULTS AND DISCUSSION

Figure 3 reports the recognition results obtained on the *Plan Restau* task in terms of word accuracy. The 95% confidence intervals are also indicated in the Figure.

The leftmost bar indicates the word accuracy obtained with the baseline speaker independent model. The second bar reports the results obtained using only the acoustic model stemming from the best matching class (determined for each utterance). The third bar shows the performance achieved with the proposed approach using model parameters from the 4 best matching classes. Finally the last bar presents the results obtained using simultaneously the model parameters from all the speaker classes.

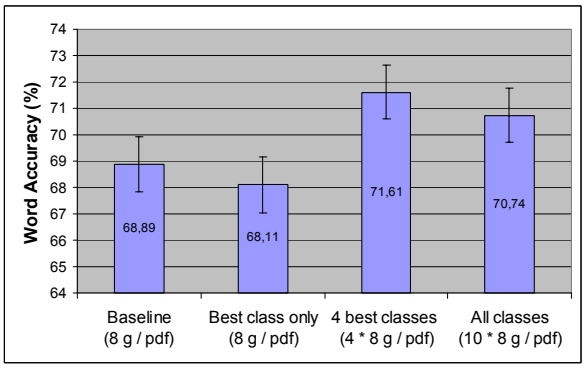


Figure 3: *Recognition accuracy on PlanResto data*.

The proposed method which selects and weight part of the acoustical space according to an estimation of the variability value provides the best recognition performance. The results are significantly better than those obtained with the baseline system and better than those obtained using only the best matching class. Using parameters from all classes together (last bar) is not as good, although the total amount of parameters is much larger.

These results show that it is beneficial to constrain the acoustic space allowed during the decoding process according to some estimation of a variability criterion, here the speaker-class. Also, the constraint must not be too strict. Using parameters stemming from several classes and adjusting the weights of the corresponding Gaussian components seems to be a good compromise.

5. CONCLUSION

A method has been proposed to handle inter-speaker variability. This method is based on combining speaker-class specific acoustic models. The pooling of the Gaussian densities for each pdf provides the global acoustic space covered by the corresponding sound for the various speakers (i.e. various variabilities). The weight coefficients are adjusted with respect to the estimated speaker-classes that best fit the speaker utterance; this limits the acoustic space that is used during decoding, and improves the recognition performance.

Although inter-speaker variability was considered here, the approach naturally extends to other kinds of variability criteria, such as speaking rate, signal to noise ratio, etc. The variability criteria can be estimated on the entire utterance as here, or on a frame by frame basis, depending on which estimation is the more relevant.

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 speech variability: a review", *Speech communication*, 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] http://www.elda.org/

[10] ETSI ES 202 212 V1.1.1 (STQ); *Distributed speech recognition; Extended advanced front-end feature extraction.*

[11] S. Krstulovic, F. Bimbot, O. Boëffard, D. Charlet, D. Fohr, O. Mella, "Optimizing the coverage of a speech database through a selection of representative speaker recordings", *Speech Communication*, vol. 48, pp.1319-1348, 2006.

[12] D. Jouvet, K. Bartkova, J. Monné, "On the Modelization of Allophones in an HMM based Speech Recognition System", *Proc. EUROSPEECH'91,* Genoa, Italy, pp. 923-926, 1991.

[13] M.D Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, J. Simonin, "Effective Human-Computer cooperative spoken dialogue : the AGS demonstrator", *Proc. ICSLP'96*, Philadelphia, PA, USA, pp. 546-549, 1996.

[14] C. Raymond, F. Béchet, N. Camelin, R. de Mori, G. Damnati, "Semantic interpretation with error correction", *Proc. ICASSP'2005,* Philadelphia, PA, USA, pp. 29-32, 2005.