Speaker Tree Generation for Model Selection in Automatic Speech Recognition

David Becerril, Oscar Saz, Carlos Vaquero, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Zaragoza, Spain
{davebv,oskarsaz,cvaquero,ortega,lleida}@unizar.es

Abstract

This paper presents the procedure and results for an automated selection of the best acoustic model for an input speaker in Automatic Speech Recognition (ASR). The procedure consists in obtaining a tree which gathers a set of representative speakers of the target population; these speakers are agglomerated by means of the Bayesian Information Criterion (BIC) until all of them are merged in the top. This tree is used when a new user accesses the system by selecting the model that best fits the speech from the speaker in order to improve the performance of the ASR system without relying on speaker dependent models trained with data from the same speaker. The results will show that the BIC metric performs correctly for building the tree, and that the selected model within the tree can outperform the whole speaker independent model in an ASR task.

Index Terms: Speech recognition, adaptation, speaker clustering.

1. Introduction

The performance of Automatic Speech Recognition (ASR) systems depends heavily on how well the acoustic models within the system match the speech characteristics from the incoming speaker. In the best situation, when some data from the speaker is available, speaker adaptation techniques allow to obtain a perfect match and the best recognition rates. Otherwise, if speaker adaptation can not be performed, a general model which covers all possible types of speech considering gender (males and females), age (children, adult or elderly), dialectal type and any other possible characteristic is used.

However, it is possible to improve the performance of a speaker independent system by using a model trained only with similar speakers; i.e. a female model to recognize a female speaker or a child model to recognize children speech. The main difficulty in this approach is to estimate in advance the speech characteristics of the user to select the best suitable model. In some cases, this information might be known a priori, but in general, an automatic procedure has to be designed to achieve this best model selection in an automated way.

The approach presented here aims to organize a set of representative training speakers into a tree which can be used when a new speaker accesses the ASR system. The tree gathers the different speakers in an agglomerative process until a top model containing all the speakers is reached. When data from a new speaker is collected, the tree is evaluated to decide which is the model which best fits the speaker so that the selected model can be used for this speaker as acoustic model in an ASR system or as seed model in a speaker adaptation stage.

This work was supported by national project TIN2008-06856-C05-04 from the Spanish government and Santander Bank scholarships

The paper is organized as follows: Section 2 will explain the proposed method of tree-based agglomerative clustering of speaker models. Section 3 will describe the proposal for selecting the most suitable model within the tree for a given input speaker. Posteriorly, Section 4 will present the experimental setup for the evaluation of the proposed methods with a set of speakers. Then, the possibilities that these techniques arise for enhanced ASR and speaker adaptation will be presented in Section 5; and, finally, Section 6 will serve as discussion for this work and will present the conclusions.

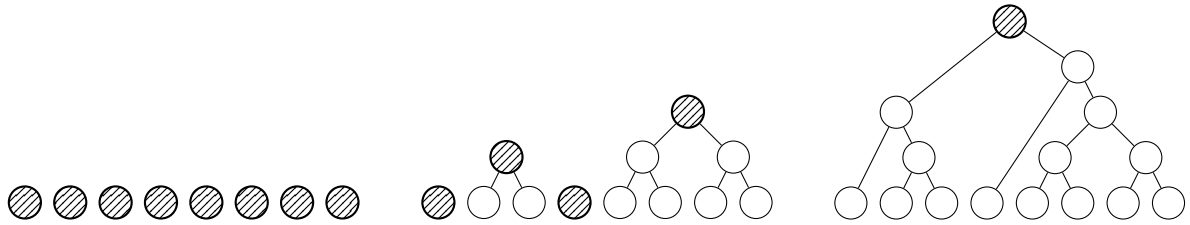
2. Automatic Speaker Tree Generation

The tree has to contain representative data from all the groups to be considered in the system. A selection of speakers has to be done in order to characterize the different groups; this may include a full set of male, female, children, and adult speakers, or a subset of them if the tree aims to model a particular type of speech.

Algorithm 1 Tree Generation

for $i = 1$ to n_{spks} **do**
  $gmm(i) =$ trainGMM()
end for
for $i = 1$ to n_{spks} **do**
  for $j = i + 1$ to n_{spks} **do**
    $distance(i, j) =$ BIC(i, j)
  end for
end for
$nodes_{rem} = n_{spks}$
repeat
  m and n so $(m, n) = \min_{i,j} distance(i, j)$
  $newnode =$ trainGMM$(m \cup n)$
  $nodes_{rem} = nodes_{rem} - 1$
  for $node \in NODES \neq newnode$ **do**
    $distance(newnode, node) =$ BIC$(newnode, node)$
  end for
until $nodes_{rem} == 1$

Each speaker (i) in the whole list of speakers (n_{spks}) is characterized by a set of N_i input speech frames $(x_i(1)...x_i(N_i))$ obtained from sufficient speech data from the speaker. A previous Voice Activity Detection (VAD) stage assures that only speech frames are fed to the system, as non-speech frames (silence, noise, etc) would lead to a poorly conditioned tree. The VAD used in the proposed system is based on the Long Term Spectral Divergence (LTSD) [1]. In order to compute the LTSD, the Long Term Spectral Envelope (LTSE) and a noise estimation have to be calculated framewise. The LTSD is calculated proportionally to the ratio between the LTSE



(a) *Initial stage* (b) *Agglomerative stages* (c) *Final tree*

Figure 1: *Example of tree generation (agglomerated nodes in white, active nodes in stripped pattern).*

and the noise estimation. Then, those frames whose LTSD value is above a certain threshold are selected, otherwise they are considered as silence and therefore discarded.

The generation of the tree is performed via an agglomerative bottom-up approach following the procedure described in Algorithm 1, with a graphical example provided in Figure 1. Initially, all the leafs in the tree (bottom nodes) are assigned to a single speaker in the corpus and are marked for agglomeration, as seen in Figure 1(a). All the speech data available from each speaker is used to train a Gaussian Mixture Model (GMM) which models the speaker and the node.

In every recursion of the tree generation algorithm, a Bayesian Information Criterion (BIC) metric is used to decide which are the pair of nodes which have to be agglomerated next. At the initial stage, the metrics between all possible pairs of nodes are calculated to perform the agglomeration of the most similar nodes. When two nodes are agglomerated, the speech data from them is used to train a GMM to model the new node and the BIC metrics between this new node and the remaining nodes are calculated. In every recursion, the number of marked nodes is decreased by one and the process of agglomerating pairs of nodes according to the BIC metric is continued (Figure 1(b)) until the top node that contains all the data from all the speakers is reached (Figure 1(c)). In the end a binary tree is obtained, as nodes are agglomerated pairwise; but the tree can be unbalanced, that is, nodes that contain more than one speaker may merge with a single-speaker node at a different tree depth, as seen in Figure 1(c).

The modeling of each node as a GMM varies depending on the depth within the tree. A fixed number of Gaussian distributions are assigned to the GMM of each bottom node. Posteriorly, when a node agglomerates two lower nodes, the complexity in the data increases. In order to compensate this effect, the number of Gaussian distributions used in the new node model equals the sum of its children's Gaussian distributions. For example, bottom nodes can be modeled as a mixture of 2 Gaussian distributions. Then, when two of these nodes are agglomerated, the resulting node will have a mixture of 4 Gaussian distributions.

When a new node is created, the GMMs of the child nodes are used as seed for the new node GMM training. For the bottom nodes, when no initial data is availabe, the GMM training is initiated using k-means labels.

2.1. BIC-based Metric

BIC criteria are generally used to determine whether two sets of data ($\underline{x} = \{x_1, x_2, \ldots, x_N\}$ and $\underline{y} = \{y_1, y_2, \ldots, y_M\}$) are more likely to be modeled into one single model (alternative hypothesis) or two separate models (null hypothesis) [2].

The decision of accepting or rejecting the alternative hypothesis of merging sequences $\underline{x}$ and $\underline{y}$ in $\underline{z} = \underline{x} \cup \underline{y}$ of length $P = N + M$ with models X and Y in a single model Z is formulated in Equation 1. Alternative hypothesis is accepted when the log-probability of the merging model is greater than the sum of the log-probabilities of the two initial models; where a parameter d represents the penalty on the merging model to compensate its higher complexity.

$$\log P(\underline{z}|Z) - \frac{1}{2} d \log(P) \geq \log P(\underline{x}|X) + \log P(\underline{y}|Y) \quad (1)$$

In our case, where the number of parameters of the merging model Z is equal to the sum of the parameters of the initial models X and Y, a valid variant of the Equation 1 is provided in Equation 2. The complexity penalty is compensated with different parameter size in both BIC hypothesis.

$$\log P(\underline{z}|Z) \geq \log P(\underline{x}|X) + \log P(\underline{y}|Y) \quad (2)$$

In the described system, the BIC-based metric is obtained reformulating the BIC decision threshold in Equation 2 to a metric formula in Equation 3. This metric is computed as the difference between the log-probabilities of the initial single models versus the merging model. The lower this $BIC(\underline{x}, \underline{y})$ is, the more probable that the two models X and Y have to be merged. High values of this metric indicates these models are likely to be different.

$$BIC(\underline{x}, \underline{y}) = (\log P(\underline{x}|X) + \log P(\underline{y}|Y)) - \log P(\underline{z}|Z) \quad (3)$$

3. Model Selection

It is now possible, when data from a new speaker is available, to select the model in the tree which best fits the speech from this incoming speaker. This selection is performed in two steps: First, the tree is pruned to a single branch, a path from the top node to one of the bottom nodes; and then, the best model from that path is chosen. Before these stages, a VAD is applied to the speech data from the speaker to discard silence in the signal, as it was done in the tree generation stage with the training speakers.

3.1. Tree Pruning

The number of nodes that the speaker tree contains is $n_{nodes} = 2 * n_{spks} - 1$. Even with a moderate number of speakers, the evaluation of all the existing models in the tree becomes a computational challenge. However, it is possible to improve this situation by pruning the tree to just one branch.

An in-depth evaluation of the tree is made with the incoming speaker and the best path is obtained evaluating its data on the tree GMM models. Starting from the top, the iterative process consist in choosing as next node-in-path the child node that maximizes the likelihood of the model X to the speaker's data $\underline{y}$: $P(\underline{y}, X)$, until a leaf is reached.

Figure 2, shows an an hypothetical pruned path, highlighted in dots.

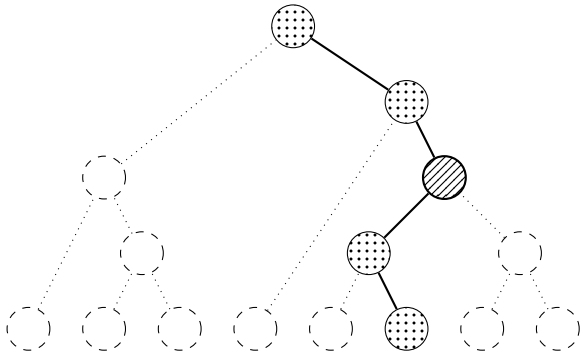


Figure 2: *Example of the selection of the best model (dotted nodes belong to the pruned tree and the stripped node is the selected node).*

3.2. Model Selection

Once the whole tree has been pruned into a single set of models, ranging from the full speaker independent model in the top to a speaker dependent model in the bottom, the best model has to be selected. This decision can be made according to different proposals: Likelihood scoring of the GMMs calculated during the training stage, likelihood scoring of the Hidden Markov Model (HMM) associated to the speech data in each node or a priori, selecting a certain node depending on its depth in the tree. In Figure 2, an hypothetically selected node is highlighted in strips.

4. Experimental Setup and Evaluation

This Section will introduce the experimental framework used in this work and the evaluation results in the speaker tree generation and lookup. The tree generation system was based in BIC criterion and GMM modeling of the nodes as explained in previous Sections. The bottom nodes (single speaker) were modeled with two GMMs, and upper nodes were the sum of the number of Gaussian distributions in their children respectively.

The feature extraction method for the tree generation was based on a standard ETSI front-end using the first 12 Mel Frequency Cepstral Coefficients (MFCC) discarding $c0$. The speech signals were windowed with a Hamming window of 25 ms length, with an overlap of 15 ms.

4.1. Experimental Corpus

The corpus used in this work was the "Alborada-I3A" corpus of disordered speech [3], that contains speech from a group of 232 unimpaired young speakers, used to build the speaker tree, and 14 young disabled speakers as potential users of the recognition system. ASR for disabled speakers is a complex task due to the many effects or their physical and cognitive impairments in their speech. For this reason, they require a proper matching to their characteristics when selecting acoustic models to be used in recognition. In our proposal, the 232 unimpaired peers will serve as reference speakers to select the best fitting model to every impaired speaker.

The group of unimpaired speakers represents the speech of individuals ranging in age 11 to 18 years old. This corpus contains one session peer speaker of the 57 words in the Registro Fonológico Inducido (RFI) [4], for a total of 13,224 isolated-word utterances.

The 14 young disabled speakers are distributed as 7 boys and 7 girls in a similar range of age (from 11 to 21 years old). Each speaker recorded 4 sessions of the RFI vocabulary, for a total of 3,192 isolated-word utterances. These speakers suffer from different developmental disorders that affect their language acquisition, resulting in a great number of mispronunciations (substitution and deletions) at the phonetic level. Physiological disorders in their vocal tract components, due to multiple physical impairments, also affect their production of speech.

4.2. Evaluating the Speaker Tree: Speaker Identification

The evaluation of the abilities of the proposed methods to build a useful speakers tree and detect correctly the best model for a given speaker was made with the following experiment. Three different trees were built with the data from the 232 reference speakers in the corpus. Each tree contained two thirds of the data (38 words) from each speaker, while the remaining third (19 words) was saved for evaluation purposes. The first tree (Set 1) comprised words 1,2,4,5,7,8... while words 3,6,9... were meant for evaluation; the second tree (Set 2) was built with words 1,3,4,6,7,9... from each speaker, keeping words 2,5,8... for evaluation; and, finally, the third tree (Set 3) had words 2,3,5,6,8,9... for generating the tree and words 1,4,7... for the evaluation.

The tree was pruned according to Section 3.1 to a single path from the top node to one of the bottom nodes. The remaining data from each speaker was evaluated through the tree as a new speaker, measuring the accuracy in which this pruned tree lead to the bottom node which was built from speech data from the same speaker.

Table 1: *Search tree accuracy.*

	Set 1	Set 2	Set 3	Average
Accuracy	79.74	79.31	81.04	80.03

The results in this task for the 3 proposed sets in Table 1 assured the method used for the tree generation and pruning as in 80.03% of the cases, the tree allowed to reach the speaker who was actually evaluating the system. In this speaker identification task, there were 232 possible competing speakers, with each speaker having just an average of 32.07 seconds of speech for training data and 16.03 seconds for the evaluation data. Only 2 Gaussian distributions formed the final nodes to model the training speaker data. This competitive performance of the tree in detecting similar speakers was encouraging in its possibilities of providing an improvement in the ASR task.

5. Use of the Speaker Tree in ASR

The purpose of the use of the tree in an ASR task is to improve the recognition rates by using an acoustic model that best matches the speech from the speaker. The experiments presented here were based on the recognition of the 14 impaired speakers on the corpus presented in the previous Section. The initial model used for the recognition was trained purely on adult speech with the 44108 noise-free signals of the adult Spanish speech databases SpeechDat-Car [5], Albayzin [6] and Domolab [7]. These data was used to train an speaker independent HMM acoustic model with a set of 744 context-dependent phonetic units and two units to model begin-end silence and interword silence; all units were modeled as 1-state units with 16 Gaussian distributions per state. 39 MFCC parameters were used for recognition, with 12 static parameters and log-energy plus their first and second derivatives. The Word Error Rate (WER) of the impaired speakers with this model was 36.69%,

showing up the big influence of the impairments of the speakers in their ability to use speech recognition; while the WER of the 232 unimpaired peers was 3.99%.

The proposal divided the data from each speaker into two subsets; one for a initial development stage and the other for evaluating the new models obtained in the development stage. A set of experiments were designed to create all the 14 different possibilities using 1, 2 and 3 sessions for the development stage and the complementary 3, 2 and 1 sessions for evaluation. The purpose was to learn how the amount of data used to tune up and adapt the system influenced the performance of the recognition stage. All the adaptation processes carried out in these experiments followed a Maximum A Posteriori (MAP) implementation [8].

Table 2: *ASR results without adapting to the speaker data.*

Development data	Model	WER
N/A	From 232 speakers	28.20%
1 session	From best node	27.41%
2 sessions	From best node	27.25%
3 sessions	From best node	27.50%

From this starting point, two cases of study were evaluated. In the first case, it was considered that the data separated for development was not transcribed, hence adaptation of the models to the speaker was not possible (although unsupervised algorithms could have been applied). If no tree were available, adaptation to the whole 232 speakers was the only possible option. The result through all the speakers with the model adapted to the unimpaired children was 28.20%, as seen in first row of Table 2. When the tree was used, the best node in the tree was estimated through the evaluation of the likelihood of the speaker data to each one of the models in the pruned tree as seen in Section 3.1. Afterwards, a new model was adapted through MAP with the data of the speakers in the selected node of the tree and this model was used in the ASR stage. The average results in this case for the three possible amounts of development data are shown in Table 2, with improvements in all cases over the system which did not use the tree information.

Table 3: *ASR results adapting to the speaker data*

Development data	Model	WER
1 sessions	From 232 speakers	19.36%
	From best node	18.96%
2 sessions	From 232 speakers	16.50%
	From best node	16.50%
3 sessions	From 232 speakers	15.29%
	From best node	14.85%

The second case of study considered that the data for the initial stage was transcribed, hence it could be used to perform supervised adaptation via MAP algorithm. Again, when no information about the tree was used, the model trained from the 232 unimpaired children was used as seed for the adaptation algorithm with the specific speakers' sessions. The results in these cases are presented in Table 3, showing the more amount of data from the speaker in adaptation, the lower the WER was. Whenever the tree was used, the best node was estimated considering the likelihood of the transcribed utterances from the speaker in a HMM-based Viterbi forced alignment, comparing the HMM models assigned to each node in the pruned tree. This model served as seed in the MAP algorithm and the results are also shown in Table 3. Again, the use of the tree implied an improvement, depending of the available training datata.

6. Discussion and Conclusions

The results presented in this work showed how the proposed method produced certain improvements in terms of WER for the ASR task. The small impact of these improvements is due to the compact group of speakers available for the tree building and recognition process (all children and young adults). Further work in tasks with sets of more differentiated speakers such as adults (especially males) and children should perform better in terms of relative improvement of the WER.

The process of generating the tree was validated by means of a speaker identification task with 232 different speakers. Different utterances from the same speakers whom built the tree, were used to obtain the most likely speaker according to the tree. The 80% of accuracy showed the ability of the algorithm to build an effective speaker tree and develop a lookup system to reach the most similar nodes for an incoming speaker.

Several aspects have appeared as result of these experiments as further work. First, other techniques can be proposed for the selection of the best node to fit the speaker's data; this may include likelihood of the GMM model or the HMM model as proposed here, as well as other techniques which may consider different issues in the algorithm, such as the amount of data that would be available for creating the HMM of each node or the depth within the tree. New tasks can also be proposed where the use of a speaker tree, like the one in this work, can produce an improved performance. Its use in speaker identification or verification task could be studied in the future, as well as the possibilities in ASR situations where very sparse data from the speaker is available to train adapted models; the selection of the best speaker independent model could help to provide better recognition in rapid adaptation systems.

7. References

[1] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detector algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[2] S. S. Chen and P. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proceedings of the 1998 ICASSP*, 1998, pp. 645–648.

[3] O. Saz, E. Lleida, C. Vaquero, and W. R. Rodríguez, "The Alborada-I3A corpus of disordered speech," in *Proceedings of the 7th LREC*, 2010, pp. 2814–2819.

[4] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*. Madrid, Spain: Ed. Cepe, 1989.

[5] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speech Dat Car. A large speech database for automotive environments," in *Proceedings of the II LREC*, Athens, Greece, June 2000.

[6] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. Mariño, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proceedings of the 1993 Interspeech)*, Berlin, Germany, September 1993, pp. 175–178.

[7] R. Justo, O. Saz, V. Guijarrubia, A. Miguel, M.-I. Torres, and E. Lleida, "Improving dialogue systems in a home automation environment," in *Proceedings of the Ambi-Sys 2008*, Québec City, Canada, February 2008.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.