Focal Speakers: a speaker selection method able to deal with heterogeneous similarity criteria

Sacha Krstulović ♣, *Frédéric Bimbot* ♣, *Delphine Charlet* ♡, *Olivier Boëffard* ♢

♣ IRISA/METISS
Campus de Beaulieu,
35042 Rennes - France
{sacha,bimbot}@irisa.fr

♡ France Télécom R&D
2, ave. Marzin
22 307 Lannion - France
delphine.charlet@francetelecom.com

♢ IRISA/CORDIAL
6 rue de Kerampont - BP 80518
22 305 Lannion Cedex - France
olivier.boeffard@univ-rennes1.fr

Abstract

In the context of the NEOLOGOS speech database creation project, we have studied several methods for the selection of representative speaker recordings. These methods operate a selection by optimizing a quality criterion defined in various speaker similarity modeling frameworks. The obtained selections can be cross-validated in the modeling frameworks which were not used for the optimization. The compared methods include K-Medians clustering, Hierarchical clustering, and a new method called the selection of Focal Speakers. Among these, only the new method is able to solve the *joint* optimization, across all the modeling frameworks, of the selection of representative speakers.

1. Presentation

The NEOLOGOS project [1] aims at creating a speech database for the French language, with the goal of answering the needs of the most recent developments in Speech/Speaker Recognition and Adaptation as well as Text-To-Speech synthesis. These recent developments promote the use of sets of specialized models instead of global models. Hence, they require some speech data distributed over a reduced number of carefully chosen representative speaker recordings, rather than distributed over a large set of non-specific speakers. Alternately, the goal of limiting the number of recorded speakers without hampering the performances of the recognition or synthesis systems meets the practical concern of reducing the database collection costs.

In this context, the corner stone lies in the speaker selection method. This method should guarantee that the subset of speakers preserves a diversity of the recorded voices, both at the segmental and supra-segmental levels. A solution to this problem relies on clustering methods.

Section 2 will expose our methodological framework, and will introduce the methods used to model the speaker similarity. Section 3 will focus on the speaker selection methods. Section 4 will comment some experimental results.

2. General framework

2.1. Approach and notations

Let M be a large number of speakers $x_i,\ i = 1, \cdots, M$, among which we want to choose a subset of $N < M$ reference speakers. In the context of the NEOLOGOS project, $N = 200$ and $M = 1000$. Let:

- $L = \{\Theta_j;\ j = 1, \cdots, N\}$ be a set of N potential reference speakers Θ_j;
- $d_A\left(x_i, \Theta_j\right)$ be a function able to measure the distance, or dissimilarity, of x_i to any reference speaker Θ_j in the modeling framework A;
- $\mathrm{ref}_A(x_i|L)$ be a function able to find out, among the list L, the reference speaker which provides the best modeling of the speaker x_i in the context of the method A:

$$\mathrm{ref}_A(x_i|L) = \arg \min_{j=1,\cdots,N} d_A(x_i, \Theta_j) \qquad (1)$$

Given the above definitions, a measure of quality can be defined for a given list L as:

$$Q_A(L) = \sum_{i=1}^{M} d_A\left(x_i, \mathrm{ref}_A(x_i|L)\right) \qquad (2)$$

This quantity measures the total cost, or total loss of quality, that occurs when replacing each of the M initial speakers by their best reference among the N reference speakers listed in L, according to the modeling method A. The smaller this total loss, the more representative the reference list. In turn, finding the optimal subset L^A of reference speakers with respect to the modeling method A translates as:

$$L^A = \arg\min Q_A(L) \qquad (3)$$

This optimization is the focus of the present paper and is detailed in section 3.

With this approach, it is also possible to evaluate a list L^A, optimized in the context of the modeling framework A, in terms of quality in the context of a different modeling framework B:

$$Q_B(L^A) = \sum_{i=1}^{M} d_B\left(x_i, \mathrm{ref}_B(x_i|L^A)\right) \qquad (4)$$

2.2. Speaker similarity modeling

Many inter-speaker metrics have been studied in the context of clustering applications (e.g., [2], [3], [4] etc). For NEOLOGOS, we have chosen to apply three methods which focus on a variety of speech modeling aspects:

– the *Canonical-Vowels* (CV) metrics intends to account for physiological differences between speakers, related to their vocal tract dimensions, in a maximum likelihood modeling framework. It implements the inter-speaker distance as a sum of Kullback-Leibler distances between mono-Gaussian models of the phonemes /a/, /i/ and /u/;

– the *Dynamic Time Warping* (DTW) metrics makes minimal modeling assumptions, provides a "direct" comparison of the

speech signals, and is affiliated with classical speech recognition techniques. It implements the inter-speaker distance as an average DTW distance between speech segments corresponding to well pronounced breath groups;

– the *Gaussian Mixture Models* (GMM) metrics makes use of the acoustic models that are employed in state-of-the-art speaker recognition. It implements the inter-speaker distance as an estimate of the Kullback-Leibler distance between the GMMs of each individual speaker [5].

These methods are described with more detail in [6].

3. Speaker selection methods

Finding a global optimum by an exhaustive search among every possible combination of speakers is not tractable in practice, due to the high number C_M^N or $\binom{M}{N}$ of possible combinations (e.g., for NEOLOGOS: $C_{1000}^{200} = \binom{1000}{200} = 6.6172 \cdot 10^{215}$). Nevertheless, this optimization problem can be understood as a clustering task. Classical clustering algorithms, able to find locally optimal solutions, include the K-Means algorithm (or a K-Medians variant) and the Hierarchical Clustering algorithm. In addition, we propose a new method called the Focal Speakers selection.

3.1. The K-Means/K-Medians algorithm

The K-Means algorithm [7] aims at grouping data in classes by locally minimizing the following criterion:

$$Q = \sum_{i=1}^{M} d(x_i, \text{ref}(x_i|C)) = \sum_{n=1}^{N} \left(\sum_{x_i \in \mathcal{C}_n} d(x_i, c_n) \right) \quad (5)$$

where C is a list of N classes $\mathcal{C}_n$ in which the data x_i will be clustered, and $c_n = \text{ref}(x_i|C)$ indicates the position of the centroid which abstracts the class $\mathcal{C}_n$. In our framework, the centroids have to be ultimately assimilated to real speakers. Besides, if this assimilation is made at each iteration, a lot of computation can be saved, because the distances between the centroids and the speakers can be read from a pre-computed matrix of inter-speaker distances. The corresponding discretized version of the K-Means algorithm, called the K-Medians, uses the following steps:

1. computation of the matrix of speaker similarities for the considered modeling method (CV, DTW or GMM);
2. random initialization, by a uniform draw of N reference speakers among the $M > N$ initial speakers;
3. assignation of each speaker to the cluster characterized by the closest reference speaker;
4. for each new cluster, determination of the reference speaker as the median speaker, i.e. the one for which the sum of the distances to every other speaker in the cluster is minimum:

$$c_n = \arg \min_{x_j \in \mathcal{C}_n} \sum_{x_i \in \mathcal{C}_n} d_A(x_i, x_j) \quad (6)$$

5. iteration of steps (3) and (4) until the N clusters stabilize.

At step 3., the assignation is done so that each of the $d(x_i, \text{ref}(x_i|C))$ terms of the sum in equation (5) diminishes or stays the same; then, at step 4., the upgrade of c_n for each class $\mathcal{C}_n$ minimizes the $\sum_{x_i \in \mathcal{C}_n} d(x_i, c_n)$ term explicitly, so that the second expression in (5) is further minimized. Therefore, the final solution will get a quality better than or equal to that of the list used for the initialization at step (2).

As a matter of fact, the result of the K-Medians is very dependent on the initialization, and the degree of quality of a locally optimal solution is undefined a priori. A solution consists in realizing a great number of runs of the algorithm, with different initializations, and to keep the local solution which reaches the best quality.

3.2. Hierarchical clustering

Whereas the K-Means/K-medians algorithm considers the data as a set of independent observations, the various versions of the Hierarchical Clustering algorithm [7] proceed by establishing a typology of the data which can be described by a tree, or *dendrogram*. In the tree, each node describes a group of observations, characteristic of a particular class of data. The building of the tree can be operated in two manners:

– *agglomerative hierarchical clustering*: the classes described in the parent nodes are determined by merging the characteristics defined in the child nodes. The nodes to merge are chosen so that they minimize the following criterion:

$$\Delta(\Theta_i, \Theta_j) = \sum_{x_k \in \pi_{i \cup j}} d_A(x_k, \Theta_{i \cup j}) - \sum_{x_k \in \pi_i} d_A(x_k, \Theta_i) - \sum_{x_k \in \pi_j} d_A(x_k, \Theta_j) \quad (7)$$

where π_j is the population of the cluster/node represented by Θ_j, and $\pi_{i \cup j}$ is the union of the π_i and π_j populations. It can be shown that this criterion corresponds to a direct optimization of the quality Q_A within the constraints of the dendrogram construction. After each merge, a new representative speaker is chosen as the centroid of the merged population.

– *divisive hierarchical clustering*: the child nodes inherit from the characteristics of their parent, but are further divided so that they refine the taxonomy of the data. The node to divide is chosen so that it minimizes the criterion (7). For each node splitting, the speaker assignations and the centroids for the two child nodes are determined by the local application of a 2-classes K-Medians on the population of the parent node. Since this K-Medians is repeated over all the parent nodes to minimize (7), the divisive version is significantly heavier than the agglomerative one.

In any case, the tree-building procedure is stopped when the number of nodes reaches the requested number of clusters (200 for NEOLOGOS). The list of reference speakers obtained by the Hierarchical Clustering procedure can be used as an initialization for the K-Medians.

3.3. The Focal Speakers method

Presentation – This method is based on empirical considerations. It starts from the hypothesis that speaker subsets with a good quality are more likely to contain some speakers of the global optimum. If this hypothesis is true, the reference speakers of the global optimum should appear more often than others in a set $\mathcal{L}_K = \{L_k;\ k = 1, .., K\}$ made of a union of locally good speaker lists. To verify this, we computed the number of occurrences of each of the M initial speakers x_i among:

(a) K random lists of N speakers;

(b) the K best lists of a great number of random lists;

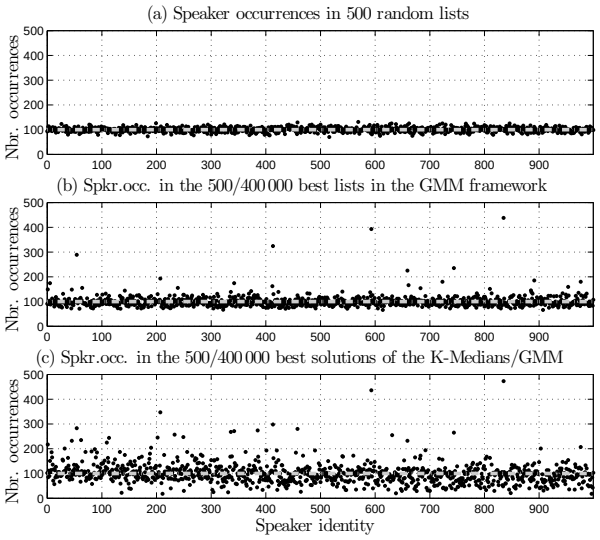

Figure 1: Number of speaker occurrences for various compositions of $\mathcal{L}_{500}$.

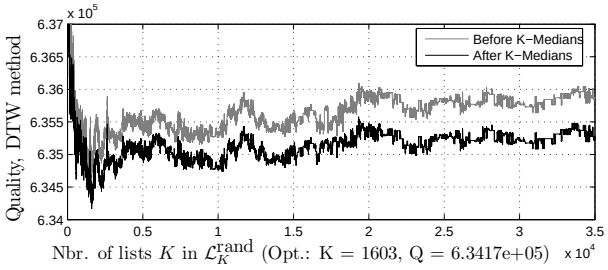

Figure 2: Quality of the lists of Focal Speakers as a function of the number K of best random lists in $\mathcal{L}_K^{\text{rand}}$, in the DTW modeling framework. (See the text for more explanations.)

(c) the K best lists among the solutions given by a great number of runs of the K-Medians.

The results are depicted in **figure 1**, for lists of $N = 200$ speakers taken among $M = 1000$ speakers, and with $\mathcal{L}_K$ gathering $K = 500$ lists taken from 400 000 initial lists. The number of occurrences of each speaker (black dots) is compared to the expected number $K \times N/M = 100$, corresponding to a uniform draw of 200 speakers among 1000. The figure shows that some speakers appear more often than the average across the series of lists characterized by their locally good quality. This suggests that there is a correlation between the quality of the lists and the fact that they contain some particular reference speakers.

Reverting this idea, we have studied if the N most frequent speakers in a set of lists characterized by their good quality would correspond to a good selection of reference speakers. Let:

- $\mathcal{L}_K = \{L_k;\ k = 1, .., K\}$ be a set of speaker lists L_k;
- $\delta_k(i) = \begin{cases} 1 & \text{if speaker } x_i \in L_k; \\ 0 & \text{else.} \end{cases}$

The number of times the speaker x_i appears in $\mathcal{L}_K$ is therefore defined as:

$$\text{Freq}(\, i \,|\mathcal{L}_K) = \sum_{L_k \in \mathcal{L}_K} \delta_k(i) \qquad (8)$$

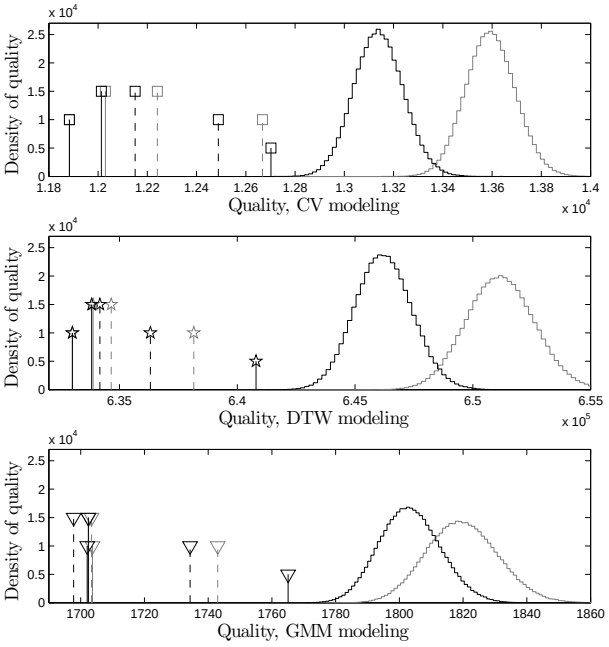

Figure 3: Evaluation of the solutions of the K-Medians, Hierarchical Clustering and Focal Speakers methods, optimized in each separate frameworks. (See the text for more explanations.)

Then, the speakers corresponding to the N highest values of Freq($i\,|\mathcal{L}_K$) can be selected to constitute a list of so called Focal Speakers. This list will be noted $L_{\text{foc}}(\mathcal{L}_K)$. Its quality $Q_A(L_{\text{foc}}(\mathcal{L}_K))$ can be computed from various sets $\mathcal{L}_K$ of "good lists", and, in particular:

– the set $\mathcal{L}_K^{\text{rand}}$ containing the K best of 400 000 random lists;

– the set $\mathcal{L}_K^{\text{kmed}}$ containing the K best of 400 000 K-Medians results.

Figure 2 illustrates the evolution of $Q_{\text{DTW}}(L_{\text{foc}}(\mathcal{L}_K^{\text{rand}}))$ versus the number K of lists in $\mathcal{L}_K^{\text{rand}}$. This evolution is shown by the gray curve. The quality of the best list in $\mathcal{L}_K$ corresponds to $K = 1$. The lower the quality value, the better the list: for every value of K, $L_{\text{foc}}(\mathcal{L}_K)$ has a better quality than the best list in $\mathcal{L}_K$. Similar results have been observed in the other modeling frameworks than DTW, as well as with $\mathcal{L}_K^{\text{kmed}}$. The most frequent speakers have been called *Focal Speakers* because they seem to concentrate the quality of the lists gathered in $\mathcal{L}_K$.

The lists of focal speakers obtained for each K can be used to initialize additional runs of K-Medians. The resulting additional gain of quality is represented by the black curve.

Joint optimization across various modeling frameworks – The Focal Speakers approach naturally suggests a joint optimization for the three speaker similarity modeling frameworks. One can search the focal speakers among a set $\mathcal{L}_{K\times\text{CV+DTW+GMM}}^{\text{rand}}$ or a set $\mathcal{L}_{K\times\text{CV+DTW+GMM}}^{\text{kmed}}$ formed by gathering the best lists obtained in CV, DTW and GMM. The corresponding results will be given in the next section.

4. Results

Figure 3 compares the solutions of the various speaker selection algorithms for each of the 3 separate modeling frameworks:

– the gray Gaussian is the density of quality of 400 000 random lists. The black Gaussian is the density of quality for the related

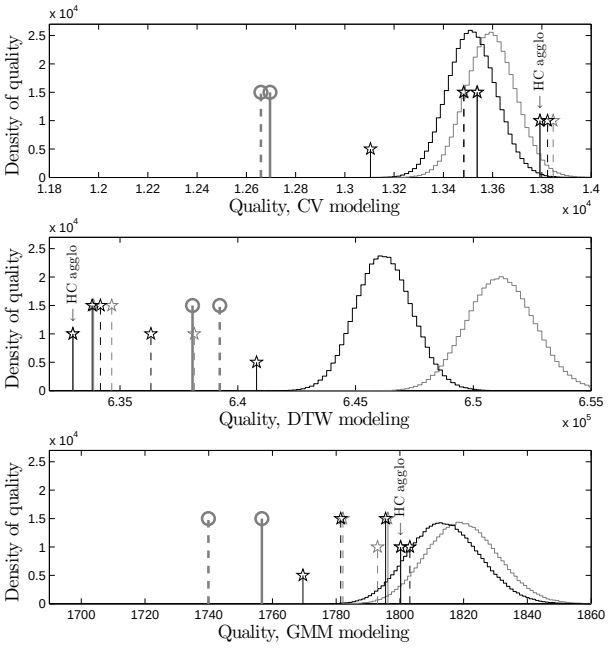

Figure 4: Solutions optimized in DTW and evaluated in the context of CV, DTW and GMM, plus solutions corresponding to the Focal Speakers optimized in the joint frameworks. (Cf. text.)

400 000 K-Medians solutions. The short flag (at level 0.5) indicates the position of the best K-Medians solution (the lower the abscissa value, the better the quality);

– the medium sized flags (at level 1) indicate the solutions of the Hierarchical Clustering (HC) in the agglomerative case (solid) and divisive case (dashed), both before (gray) and after (black) an additional run of K-Medians;

– the taller flags (at level 1.5) indicate the solutions of the Focal Speakers method for the optimal K in $\mathcal{L}_K^{\text{rand}}$ (dashed) and $\mathcal{L}_K^{\text{kmed}}$ (solid); these solutions are indicated before (gray) and after (black) an additional run of K-Medians.

The agglomerative HC reaches better results than the divisive HC. The Focal Speakers method reaches qualities comparable to the agglomerative HC, both with the best random lists and with the best K-Medians.

Figure 4 considers the solutions optimized in the DTW framework, and evaluates them in the context of the alternate modeling methods. It shows that an optimal quality in a given modeling framework does not necessarily guarantee a good quality in the other ones. For example, the solution brought by the agglomerative HC applied in the DTW framework (distinctly marked in the figure) is the best in its framework of origin, but has a low quality with respect to CV and GMM modeling. Similar effects have been observed in the other cases of match or mismatch between the optimization context and the evaluation context.

The bold gray flags indicate the quality of $L_{\text{foc}}(\mathcal{L}^{\text{rand}}_{K\times\text{CV+DTW+GMM}})$ (dashed) and of $L_{\text{foc}}(\mathcal{L}^{\text{kmed}}_{K\times\text{CV+DTW+GMM}})$ (solid) for $K = 500$ lists in each framework (1500 lists in total). As opposed to the previous case, the quality of these optimal lists of reference speakers is consistently good in all the frameworks. Besides, these lists can be used to initialize an additional K-medians in each of the separate frameworks, giving 3 more lists with an even better (or same) quality in all the frameworks (not depicted).

Using $K = 500$ lists is, for the moment, an ad-hoc choice. We have observed that it does not influence so much the quality of the result: taking the 1000 best lists of each framework gave comparable results. Nevertheless, more elaborate ways to compose $\mathcal{L}_{\text{CV+DTW+GMM}}$ could be studied.

5. Conclusions and perspectives

In the context of the constitution of the NEOLOGOS speech database, we have compared several methods for the selection of representative speaker recordings: one based on K-Medians clustering, one based on Hierarchical clustering and one based on a new method called the Focal Speakers selection. These methods can find some optimal selections of speakers in a variety of speaker similarity modeling contexts. Besides, the optimal solutions that they provide can be cross-validated in the modeling contexts which were not used for the optimization. While the Hierarchical Clustering gives the best selection for each isolated modeling framework, only the new Focal Speakers method is able to reach solutions having a consistently good quality across all the modeling frameworks. We believe that this new approach deserves to be further investigated and tested in other contexts, since it offers an interesting strategy for selecting optimal subsets of data across multiple representations and quality criteria.

6. Acknowledgements

This work was partially funded by the French Ministry of Research in the framework of the TECHNOLANGUE program.

7. References

[1] E. Pinto, D. Charlet, H. François, D. Mostefa, O. Boëffard, D. Fohr, O. Mella, F. Bimbot, K. Choukri, Y. Philip, and F. Charpentier, "Development of new telephone speech databases for French: the NEOLOGOS Project," in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04*, 2004.

[2] M. Padmanabhan, L. Bahl, D. Nahamoo, and M. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition system," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 71–77, 1998.

[3] S. Johnson and P. Woodland, "Speaker clustering using direct maximisation of the MLLR-adapted likelihood," in *ICSLP*, vol. 5, 98, pp. 1775–1779.

[4] M. Naito, L. Deng, and Y. Sagisaka, "Speaker clustering for speech recognition using vocal tract parameters," *Speech Communication*, vol. 36, no. 3-4, pp. 305–315, 2002.

[5] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in Automatic Speaker Verification using Kullback-Leibler distances," in *Proc. ICASSP 2002*, May 2002.

[6] D. Charlet, S. Krstulović, F. Bimbot, O. Boëffard, D. Fohr, O. Mella, F. Korkmazsky, D. Mostefa, K. Choukri, and A. Vallée, "Neologos: an optimized database for the development of new speech processing algorithms," in *Proc. Interspeech'05*, September 2005.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: John Wiley and Sons, 2001.