# Neologos: an optimized database for the development of new speech processing algorithms

*Delphine Charlet* ♣, *Sacha Krstulović* ♢, *Frédéric Bimbot* ♢, *Olivier Boëffard* ♠,*Dominique Fohr* ♡, *Odile Mella* ♡, *Filip Korkmazsky* ♡, *Djamel Mostefa* †, *Khalid Choukri* †, *Arnaud Vallée* ‡

♣France Télécom R&D, 2,av. Marzin 22307 Lannion, France delphine.charlet@francetelecom.com
♢IRISA, Campus de Beaulieu 35042 Rennes, France {sacha,bimbot}@irisa.fr
♠IRISA, 6 rue de Kerampont 22300 Lannion, France olivier.boeffard@univ-rennes1.fr
♡LORIA, Campus Universitaire BP239 54506 Vandoeuvre Cedex, France {dominique.fohr,odile.mella}@loria.fr
†ELDA,55-57 rue Brillat-Savarin 75013 Paris, France {choukri,mostefa}@elda.org
‡TELISMA, 9 rue Blaise Pascal, 22300 Lannion, France avallee@telisma.com

## Abstract

The Neologos project is a speech database creation project for the French language, resulting from a collaboration between universities and industrial companies and supported by the French Ministry of Research. The goal of Neologos is to re-think the design of the speech databases in order to enable the development of new algorithms in the field of speech processing. A general method is proposed to optimize the database contents in terms of diversity of the recorded voices, while reducing the number of recorded speakers.

## 1. Presentation

### 1.1. General goals

The state of the art techniques in the various domains of Automatic Speech Processing (be it for Automatic Speaker Recognition, Automatic Speech Recognition or Text-To-Speech Synthesis by computers) make extensive use of speech databases. Nevertheless, the problem of the optimization of the contents of these databases with respect to the requested task has seldom been studied [1]. The usual definition of speech databases consists in collecting a volume of data that is supposed sufficiently large to represent a wide range of speakers and a wide range of acoustic conditions [2, 3]. Nevertheless, identifying and omitting some redundant data may prove more efficient with respect to the development and evaluation costs as well as with respect to the performances of the targeted system [1]. Alternately, the most recently developed speech recognition and adaptation algorithms tend to make use of several specialized models instead of a unique general model, and hence require an important volume of data to guarantee that the variability of speech will be accurately modeled. Similarly, the most recent advances in Text-To-Speech synthesis (TTS) require the availability of a wider range of speakers to investigate the degradation of quality which is still noticeable in the synthetic voices. Hence, the above-mentioned developments require a much larger quantity of data per speaker than traditional databases can offer. Nevertheless, the increase in the collection cost for such newer and larger databases should be limited as much as possible.

Thus the NEOLOGOS project focuses on optimizing the contents of the speech databases in order to obtain a guarantee on the diversity of the recorded voices, both at the segmental and supra-segmental levels. In addition to this scientific objective, it addresses the practical concern of reducing the collection costs for new speech databases.

### 1.2. Context of the Neologos project

The starting point of this work is to consider that the variability of speech can be decomposed along two axes, one of speaker-dependent variability and one of purely phonetic variability. The classical speech databases [3] seek to provide a sufficient sampling of both variabilities by collecting few data over many random speakers (typically, several thousands). Conversely, Neologos proposes to optimize explicitly the coverage in terms of speaker variability, prior to extending the phonetic coverage by collecting a lot of data over a reduced number of *reference speakers*.

In this framework, the reference speakers should come out of a selection process which guarantees that their recorded voices are non-redundant but keep a balanced coverage of the voice space. Thus, the collection of the Neologos corpus is a three stage process:

1. the BOOTSTRAP database is collected by recording a first set of 1,000 different speakers over the fixed telephone network. The recorded utterances are a set of 45 phonetically balanced sentences, identical for all the speakers and recorded in one call. Such sentences are optimized to facilitate the comparison of the speaker characteristics;
2. a subset of 200 reference speakers is selected through a clustering of the voice characteristics of the 1,000 bootstrap speakers.
3. the final database of 200 reference speakers, called IDIOLOGOS, is collected. The reference speakers are requested to pronounce a large corpus of 450 specific sentences, identical for all the speakers, in 10 successive telephone calls that must be completed in a short period of time to avoid shifts in the voice characteristics.

This paper focuses on the second stage of the process: the extraction of the reference speakers. This task has been interpreted as a *clustering task*, which consists in partitioning the voice space in homogeneous subspaces that can be abstracted by a single reference speaker. We formulate this problem in a general framework which remains compatible with a variety of speech/speaker modeling methods, across which some lists of reference speakers can be compared and jointly optimized.

Section 2 exposes our speaker selection methodology and the design of the related corpus. Section 3 proposes and discusses some particular instances of speaker similarity metrics. Section 4 presents the clustering method. Section 5 exposes some experimental results, while section 6 discusses some conclusions and perspectives.

## 2. Methodology and corpus

### 2.1. Formulation of the approach, notations

*2.1.1. Reference speakers*

Let $M$ be a large number of speakers $x_i,\ i = 1, \cdots, M$, among which we want to choose $N < M$ reference speakers. Let $L = \left\{\Theta_j^A;\ j = 1, \cdots, N\right\}$ be a given set of $N$ speaker prototypes $\Theta_j^A$. The prototypes can be understood either as models of some sets of speakers, or as models of a single, observed speaker. They depend on a modeling paradigm $A$. Let $d_A\left(x_i, \Theta_j^A\right)$ be a function able to measure the distance, or dissimilarity, of $x_i$ to any prototype $\Theta_j^A$ in the modeling framework $A$. The lower the distance, the better $\Theta_j^A$ models $x_i$.

Let $\text{ref}_A(x_i|L)$ be a function able to find out, among the list $L$, the prototype which provides the best modeling of the speaker $x_i$ according to the method $A$. Given the above definitions, it can be obtained as:

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

If each of the prototypes $\Theta_j^A$ refers to a unique speaker, interpreting $\text{ref}_A(x_i|L)$ as the identity of a reference speaker is straightforward. Conversely, if each of the $\Theta_j^A$ refers to a set of speakers (e.g., if the $\Theta_j^A$ are models based on some pooled speaker data), then an additional step is needed to relate $\text{ref}_A(x_i|L)$ to a unique speaker identity.

*2.1.2. Quality of a list of reference speakers*

Given the ability to represent every speaker $x_i$ of the initial set by a reference speaker issued from a given list $L$, then:

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

measures the total cost, or total loss of quality, that occurs when replacing each of the $M$ initial speakers by their best prototype among the $N$ models listed in $L$, according to the modeling method $A$. The smaller this total loss, the more representative the reference list.

*2.1.3. Optimal list of reference speakers*

In turn, finding the optimal list $L^A$ of reference speakers with respect to the modeling method $A$ translates as:

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

Due to the dimensions of the databases, solving this optimization problem by an exhaustive search across all the possible combinations of $N$ speakers taken among $M$ speakers is infeasible due to the huge number of combinations $C_M^N = \binom{M}{N} = \frac{M!}{N!(M-N)!}$. Nevertheless, it is possible to use heuristic methods such as Hierarchical Clustering or K-means to find locally optimal solutions.

*2.1.4. Comparison of reference lists*

Within equation (2), the quality of any reference list $L$ can be measured. In particular, $L$ can be a list $L^B$ issued from an optimization in the modeling framework $B$:

$$L^B = \left\{\Theta_j^B;\ j = 1, \cdots, N\right\} = \arg\min Q_B(L) \qquad (4)$$

In this case, the reference speakers can be attributed from $L^B$ *with respect to an alternate modeling framework* $A$:

$$\text{ref}_A(x_i|L^B) = \arg \min_{j=1,\cdots,N} d_A(x_i, \Theta_j^B) \qquad (5)$$

It follows that the quality of a selection of reference speakers $L^B$ made in the framework of the modeling method $B$ can be evaluated in the scope of the modeling method $A$:

$$Q_A(L^B) = \sum_{i=1}^{M} d_A(x_i, \text{ref}_A(x_i|L^B)) \qquad (6)$$

This case illustrates the fact that the quality defined by equation (2) brings a general answer to the problem of comparing some reference lists, even when the lists come from different modeling frameworks. With this definition, it is possible to evaluate if a selection of reference speakers made with respect to the modeling method $A$ is “good” in the scope of the modeling method $B$. Defining the similarity of the lists in the space of the qualities is more general than trying to implement a direct comparison of the lists’ contents.

*2.1.5. Calibration of the measure of quality*

For the quality of a reference speaker selection to be interpretable and comparable across several modeling criteria, it is necessary to *calibrate* it. This is done by ranking $Q_A$ with respect to an estimate of the distribution of qualities, estimated from a “big enough” number of randomly generated lists of reference speakers. In a non-parametric framework, the values of $Q_A\left(\mathcal{L}^{\text{rand}}\right)$ are simply sorted in decreasing order, i.e., from the worst random list to the best. To evaluate a particular list $L$, we rank $Q_A(L)$ against the sorted qualities and divide the result by the total number of random lists. This normalized rank is called a Figure Of Merit (FOM). It is very easily interpretable: $\text{FOM}_A(L) = 80\%$ means that the list $L$ is better, in the framework of $A$, than 80% of the random lists in $\mathcal{L}^{\text{rand}}$. The closer to 100%, the better the list.

### 2.2. Corpus design and collection

*2.2.1. Repartition of the speakers*

The BOOTSTRAP database is balanced across gender, age and regional characteristics. Enhancements with respect to existing French databases such as SpeechDat [4] include a finer repartition in terms of geographic area (twelve distinct French regions are used), as well as a better representation of elderly speakers (60 and more, with a proportion approximately equal to that of the three other age ranges).

*2.2.2. Linguistic contents and phonetic alignment*

The corpora are constructed by processing sentences from large publicly available newspaper corpora in French. Automatic corpora reduction methods [5] are used to extract a subset of sentences meeting a criterion of minimal representation of all the phonemes, as well as a criterion of minimal representation of diphone classes. A phonetic alignment has been obtained by matching the corresponding orthographic transcriptions to the spoken utterances, with the help of a HMM-based labeling tool [6].

# 3. Modeling the speaker similarity

### 3.1. Speaker similarity

As seen in section 2, our method is based on the definition of a distance $d_A(x_i, \Theta_j^A)$ between a speaker $x_i$ and a cluster model $\Theta_j^A$ within a modeling framework $A$. In the case where the prototypes $\Theta_j^A$ can be abstracted by individual speakers $\hat{x}(\Theta_j^A)$ (possibly, the centroid of the cluster), this distance can be understood as an explicit speaker similarity $d_A(x_i, x_j)$, measured between two speakers $x_i$ and $x_j$ via a modeling method $A$. Many inter-speaker metrics have already been studied in the context of some clustering applications (e.g. [7], [8], etc). These metrics reflect a diversity of aspects of speech modeling. As our method enables considering various criteria, we have considered a panel of four methods which focus on a variety of speech modeling aspects: Canonical-Vowels (CV), Dynamic Time Warping (DTW), Gaussian Mixture Models (GMM) and HMM affiliated phonemes models (HMM). Each of the corresponding metrics is detailed in the following sections. All metrics are implemented with MFCC features.

### 3.2. Gaussian models of Canonical Vowels

This metrics accounts for physiological differences between speakers, related to their vocal tract dimensions, in a maximum likelihood modeling framework. We have more particularly considered the three cardinal vowels $/a/$, $/i/$ and $/u/$, located at the extremes of the vocalic triangle, because their spectral characteristics are directly related to the shape of the vocal tract.

For each phoneme $\alpha = /a/, /i/, /u/$, and denoting by $p_i^\alpha$ the Gaussian model of the phoneme $\alpha$ for speaker $x_i$, the similarity metrics between speakers $x_i$ and $x_j$ with respect to $\alpha$ is defined as:

$$d_\alpha(x_i, x_j) = KL(p_i^\alpha || p_j^\alpha) + KL(p_j^\alpha || p_i^\alpha) \qquad (7)$$

where $KL$ denotes the Kullback-Leibler divergence. A global distance $d_{\text{CV}}$ can be defined as a simple sum of the phoneme-dependent distances:

$$d_{\text{CV}}(x_i, x_j) = d_{/a/}(x_i, x_j) + d_{/i/}(x_i, x_j) + d_{/u/}(x_i, x_j) \quad (8)$$

### 3.3. A DTW-based metrics

Comparing two pronunciations of the same sentence by two different speakers through Dynamic Time Warping (DTW) amounts to computing a distance which makes only minimal modeling assumptions, stays very close to the original signal, and is affiliated with classical speech recognition techniques. In our framework, the DTW distance is computed between breath groups, which represent portions of signal which are long enough to account for various large scale speech variability phenomena (e.g., co-articulation, utterance speed etc.) while staying quite homogeneous. They have been manually determined by a phonetician expert. 160 breath groups have been obtained from the 45 reference sentences. They have an average length of 900 ms.

For a pair of speakers $(x_i, x_j)$, the DTW distance is considered only between the correct pronunciations of the breath group for both speakers. In practice, for any pair $(x_i, x_j)$ of speakers, about 150 of the 160 possible breath groups are correctly pronounced. The total distance between the two speakers is given by the average DTW distance over these pronunciations. Given the displacement constraints used in the DTW, this distance is symmetrical.

### 3.4. GMM-based speaker modeling

The Gaussian Mixture Models (GMMs) are the basis of the state of the art in the domain of Automatic Speaker Recognition [9]. In this framework, speaker dependent Gaussian Mixture Models (GMMs) are trained on the phonetically balanced sentences for each speaker of the bootstrap database. A speaker similarity metrics is defined as an estimate of the Kullback-Leibler divergence between such models, through a Monte-Carlo method [10].

### 3.5. HMM-based modeling

In this framework, some phoneme models are trained as the states of Hidden Markov models. To ensure that there is enough data for each model, they are based on pooled data sets comprising several speakers. The prototypes $\Theta_j^{\text{HMM}}$ are models corresponding to pools $\pi_j$ of speakers and they are a result of a hierarchical clustering for building phone models, in a maximum likelihood framework. As prototypes refer to pools of speakers, the speaker similarity measure is defined via a degree of similarity to the abstract models $\Theta_j^{\text{HMM}}$, instead of being established directly between the speakers. For each of the abstract prototypes, a reference speaker can be chosen as the member of the pool which is the most similar to the whole model:

$$\hat{x}\left(\Theta_j^{\text{HMM}}\right) = \arg\max_{x_k \in \pi_j} \mathcal{L}_k\left(\Theta_j^{\text{HMM}}\right) \qquad (9)$$

# 4. Speaker selection combining various criteria

According to the methodology exposed in section 2, the list of reference speakers is found by minimizing the quality criterion defined by the equation 3. This is done separately in the various modeling frameworks which define an inter-speaker metrics (CV, DTW and GMM). Three optimization methods have been applied, based on heuristic considerations: (a) a modified version of the K-Means algorithm where the *mean* of the algorithm is replaced with the *median* (since the centroid must correspond to an actual speaker instead of a virtual averaged speaker), (b) a Hierarchical Clustering algorithm, with an agglomerative and a divisive version, and (c) a new method, called the Focal Speakers selection, which showed good experimental results for this problem. These methods are extensively described and studied in [11].

The solutions issued from the various speaker selection algorithms can be evaluated and ranked across the different similarity modeling methods with the help of the FOM defined in section 2.1.5. As a matter of fact, the final list can be choosen as the one with the best average FOM over the three modeling criteria, given an additional minimal boundary of the FOM within each criterion.

# 5. Results

We have been able to extract several lists of reference speakers reaching good scores in the above-defined selection process (i.e., having a FOM=100 for each and every of the CV, DTW and GMM criteria). By defining some additional coverage consideration, the clustering method based on the HMM phone models has helped us to determine the final list to be recorded for the IDIOLOGOS database. The collection of this database is an ongoing process.

An interesting analysis is then to rebuild clusters with the 1000 speakers of the bootstrap database, around the 200 speakers of the Idiologos database, for each metrics separately. Here, a cluster is built with all the speakers who share the same reference speaker as defined in equation 1. The distribution of the size of the clusters for each metrics is plotted in figure 1. By definition, the average size of the clusters is 1000/200=5. The CV metrics is the metrics which leads to the most uniform distribution, compared to the GMM metrics which has the highest numbers of isolated speakers: for GMM, 121 speakers out of 1000 are in clusters of size one, whereas, for the CV metrics, only 48/1000 speakers are isolated.

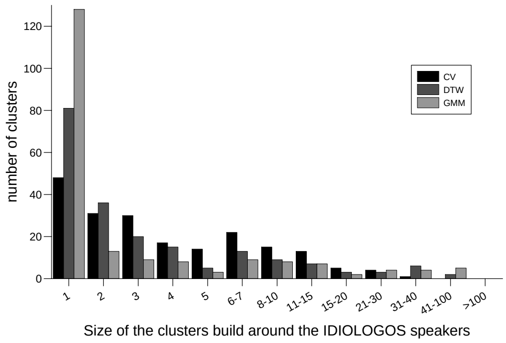

Figure 1: Distribution of the size of the clusters for each metrics

We have compared BOOTSTRAP and IDIOLOGOS in terms of age/gender/accents distribution (criteria which were not explicitly used in the extracting process). The gender distribution is the same for both databases, the accent distribution is slightly modified. The age distribution is plotted in figure 2. It is modified by the extraction process, emphasizing the contribution of the elderly people.

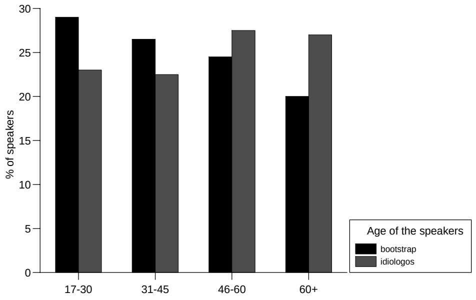

Figure 2: Age distribution

## 6. Conclusions and perspectives

We propose a method to optimize a speech database through the selection of reference speaker recordings. The optimization aims at keeping a diversity of voices while pruning the number of speakers. Hence, it is based on the notion of a speaker (dis-)similarity metrics, and of a measure of quality for some lists of reference speakers. The quality corresponds to the capacity of a reference list to keep a lot of similarity with the pruned speakers. Our implementation of this paradigm proposes, but is not limited to, 4 different ways to model the speaker dissimilarity. Then, we propose to determine the lists of reference speakers through a local optimization based on some clustering methods.

This work represents the foundation of a new framework for the optimization of speech databases. The proposed method is flexible and open to the use of other measures of speaker dissimilarity or other quality optimization schemes.

The collection of the complementary database for the selected reference speakers is an ongoing task done by our partners in the project, TELISMA and ELDA . It will be distributed by ELDA. Further work will consist in evaluating a posteriori the modeling capabilities of the IDIOLOGOS database, issued from the selection of reference speaker, compared with the modeling capabilities of the usual speech database, for instance in the framework of speech recognition.

## 7. Acknowledgements

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

## 8. References

[1] A. Nagorski, L. Boves, and Steeneken, “Optimal selection of speech data for automatic speech recognition systems,” in *ICSLP*, 2002, pp. 2473–2476.

[2] R. Lippmann, “Speech recognition by machines and humans,” *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[3] D. Iskra and T. Toto, “Speecon - speech databases for consumer devices: Database specification and validation,” in *LREC*, 2002, pp. 329–333.

[4] ELDA, 2005, see: http://www.elda.org/ for the specifications of the currently available SpeechDat databases.

[5] H. François and O. Boëffard, “Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem,” in *Proc.Eurospeech’01*, 2001.

[6] O. Mella and D. Fohr, “Two tools for semi-automatic phonetic labeling of large corpora,” in $1^{st}$ *International Conference on Language Resources and Evaluation*, may 1998.

[7] 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.

[8] 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.

[9] A. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] 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.

[11] S. Krstulovic, F. Bimbot, D. Charlet, and O. Boeffard, “Focal speakers: a speaker selection method able to deal with heterogeneous similarity criteria,” in *Interspeech’05*, 2005.