Divergence of Languages among Brazilian Indigenous Peoples

Lianet Sepúlveda-Torres*, Carlos D. Maciel, José C. Pereira, Alan P. Pinheiro.
School of Engineering of São Carlos
São Paulo, Brazil
*lianet@sc.usp.br

Paulo R. Scalassara
Institute of Physics of São Carlos
São Paulo, Brazil

Abstract— This paper presents a study of the divergence of languages among Brazilian indigenous peoples using a dendogram to represent the differences quantified by means of the KL divergence among the PDF estimations of the samples. To estimate the PDF, a method based on a mixture of Gaussians was used and compared with the common technique of normalized histogram estimation. Initial tests were performed with only ten signals of samples from different languages and based on the results; three main groups of languages could be identified.

Keywords; KL divergence; Gaussian mixture models; dendogram; indigenous languages

I. INTRODUCTION.

Linguistic studies have shown that some Brazilian indigenous people's languages are more alike than others, showing mutual influences, common origins and diversification processes that have occurred over time. Knowing the entire language repertoire and their relations has been a challenge to linguists. However, by considering common origins, linguists can create families of languages, establishing similarities among them.

This paper deals is the representation of a hierarchical structure for words belonging to the main Brazilian indigenous people's languages. Therefore it is desirable to develop an algorithm that assists the comparison of different words spoken by the indigenous peoples.

In applications dedicated to speech signal processing it is common to study the probability density function (PDF) of the signals [1], producing a compact representation that depends on a small number of parameters and enables the use of similarity measurements to compare this type of signals.

There are basically three approaches for estimating the PDF of a signal: parametric, semi-parametric and nonparametric [2]. A popular technique is the Gaussian Mixture Model (GMM), which is widely used as it is a more powerful tool when compared to nonparametric estimators that can only estimate a family of density functions [3, 4]. In this study, for comparison, the histogram technique was used.

In the literature, there exist several ways to estimate the similarities of two models based on the PDF. One of the most used is the relative entropy, also known as the Kullback Leibler divergence (KL) [5]. It is an asymmetric divergence that measures separation or disparities between two PDFs. This divergence is used in many aspects of speech and image recognition [6]. Following the method in [6], it is necessary to modify the KL divergence in order to construct a distance measurement between the PDF models. The use of the KL divergence has been widespread in many fields of science; therefore its application to GMM is very natural and frequent in the fields of speech and image recognition [6].

Our proposal is to compare words belonging to ten indigenous communities, with the use of KL divergence modified to infer a similarity structure among the words. Therefore the estimation of the PDF of the speech signals using GMM was proposed as a first approach. This preliminary model will be used to represent the signal. Next, the KL divergence is used to measure the dissimilarity between two voice models. Finally, a dendogram is generated based on the modified KL divergence, showing the similarities between speech signals. The dendogram is a representation based on a symmetric distance, but the KL divergence is asymmetric, therefore a symmetrization becomes necessary. In [7, 8] several methods are presented to symmetrize the KL divergence.

II. THEORY

This section presents a description of the PDF estimate based on GMM and the KL estimate with PDF based on GMM. The methods for symmetrizing the KL divergence are also presented, as a symmetric measurement is necessary to establish a comparison between two PDFs.

A. Gaussian Mixture Models

The speech signal, as a time series, is represented by a vector of N points and its PDF, as a mixture of Gaussian functions, and has the following form:

$$f(x,\Theta)=\sum_{k=1}^{K} p_k g(x,m_k,\sigma_k) \qquad (1)$$

where $f(x,\Theta)$ represents the mixture and

$$g(x, m_k, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-m_k}{\sigma_k})^2} \quad (2)$$

is the $K-th$ Gaussian function of the mixture, where $\Theta = [\theta_1,...,\theta_k] = [(p_1, m_1, \sigma_1),...,(p_k, m_k, \sigma_k)]$ is a vector of length K containing the mixing probabilities p_k, mean m_k and standard deviations σ_k of the K Gaussian functions [9].

For the estimation of the GMM parameters, the Expectation Maximization algorithm (EM) was used, as it is a frequently used technique to obtain the PDF in both univariate and multivariate cases [10]. Also, the EM algorithm has become a popular tool in statistical estimation problems involving incomplete data or in problems which can be posed in a similar form, such as mixture estimation [11, 12].

Given an initial parameters set for each Gaussian function of the mixture and the speech vector, the EM algorithm is implemented to estimate the parameters of the Gaussian functions. In an iterative form, the algorithm computes the Maximum Likelihood to estimate the parameters of the models, for which the input data are the most likely [13].

The EM algorithm has two principal processes for each iteration: E-step and M-step. In the E-step, the approximation PDF is estimated. This estimation is the new input for future iterations. In the M-step, the likelihood function is maximized according to the parameters estimated in the E-step. The likelihood increases in each iteration, guaranteeing the convergence of the method [13].

B. Histogram

The PDF is also estimated using the histogram method, to evaluate the PDF estimative based on GMM. For this histogram estimation, a computational approach with a fixed number of bins was proposed. The algorithm receives the number of bins and the analyzed signal as input, and the frequencies of the samples are obtained. The histogram is normalized by the sum of the occurrences to obtain unitary probability.

C. Kullback-Leibleir Divergence

Considering that $f(x)$ and $g(x)$ are probability density functions estimated using GMM, the KL divergence between them, that is, a non-negative function and equal to zero only when $f(x) = g(x)$, can be defined as

$$KL(f \parallel g) = \sum f(x) \log \frac{f(x)}{g(x)} \quad (3)$$

The KL divergence can be easily obtained, but there is no closed-form expression when it is estimated for two GMMs [6,14]. In [6], the authors discuss several techniques to estimate the KL between two Gaussian mixtures. They present some methods used to replace the KL divergence with other functions that can be efficiently computed. One of those is the Monte Carlo method, which is commonly used to compute this divergence.

Monte Carlo technique may cause a significant increase in the computational complexity and the KL non-negative property does not hold [6]. This paper proposes a computational method to estimate the KL divergence between two GMMs. Such an approach approximates the KL between two Gaussian mixtures using the envelope of the mixture.

After estimating the GMM, which represents the PDF of the analyzed signal, the envelope of the mixture is obtained and used to estimate the KL divergence. Vectors $f(x)$ and $g(x)$ are divided into a rectangular grid with i equal cells with fixed Δx. In each cell, occurrences $f(x_i)$ and $g(x_i)$ are considered for $f(x)$ and $g(x)$, respectively. Equation (3) can, then, be re-written as

$$KL(f \parallel g) \approx \sum_{i=0}^{N} f(x_i) \log \frac{f(x_i)}{g(x_i)} \quad (4)$$

where N is the total number of cells. Tests showed that the proposed computational method satisfies the three properties of the KL divergence.

Several functions are presented in the literature to symmetrize the KL divergence. In [7, 8], various solutions that can be considered for the KL symmetry problem are presented. In [7], the operations used for symmetrizing are the average geometric and harmonic means, which are shown to be equivalent. We adopted the average mean, which can be calculated using the equation

$$KL_s(f \parallel g) = \eta KL(f \parallel g) + (1-\eta) KL(g \parallel f) \quad (5)$$

Consider that $\eta = 1/2$, it reduces to

$$KL_s(f \parallel g) = \frac{KL(f \parallel g) + KL(g \parallel f)}{2} \quad (6)$$

In this approach, given a speech representation by a PDF (estimated by a GMM), a similarity measurement is

[1] http://www.gnu.org/software/gsl/
[2] http://www.gnuplot.info/

defined between two speech signals using the KL divergence.

III. THE PROPOSED ALGORITHM

The algorithm was developed using C++ with GSL[1] and Gnuplot[2] libraries and designed using object-oriented programming.

The speech signals used in this study were borrowed from the Brazilian Indian Museum[3]. The museum's database stores the basic vocabularies of the languages from the ten main Brazilian indigenous communities. There are several words spoken by male people of each isolated community. These signals were recorded in mono-channel WAV format to preserve the fidelity of the speech signal with sampling frequency of 22.050 Hz. As the duration of each signal is different, the signals have various sizes.

To organize the functionalities, the algorithm is separated into three main steps. The first step is the pre-processing module, where the signal is normalized. This normalization consists in removing the start and ending of the speech signals due to the presence of silence, using an established threshold. Also, the amplitude of the signals is normalized to levels between -1 and 1. The envelope of the signal is then extracted using the Hilbert transform, eliminating the fast oscillations [16]. The resulting signal has amplitudes that range from 0 to 1 and is used in the next step to obtain the PDF.

The second step of the algorithm estimates the PDF using GMM. This is performed by an EM algorithm, according to [9]. The algorithm requires both the number of Gaussian components of the mixture and the pre-processing speech signals as input. The same number of Gaussian components was used for all PDF estimations. The initial parameters of the Gaussian functions are randomly initialized. The output of this module is a PDF based on Gaussian mixture for each input signal.

After the PDF estimations, the KL divergence matrix is calculated. This matrix is square with main diagonal equal to zero, but it is non-symmetrical. In order to obtain the dendogram of Fig 3, which shows the similarities among the GMM, the matrix was symmetrized. Such a procedure was performed using the assumptions in [7].

IV. RESULTS

This section presents the main results of the proposed method. Ten speech signals representing the word "water" were selected for the tests.

As previously shown, a crucial step of this algorithm is the PDF estimation using GMM, as the calculation of the similarities depends on it. In order to verify the accuracy and reliability of the method, a comparison with the PDF estimated using histograms was made, as mentioned in Section II.

Fig 1 illustrates the normalized histogram and mixture PDF envelope of one example of speech signals. The signal corresponds to the word "water" spoken by a male person from the *Kamayura* community. To evaluate the differences between the two PDF estimates, the mean

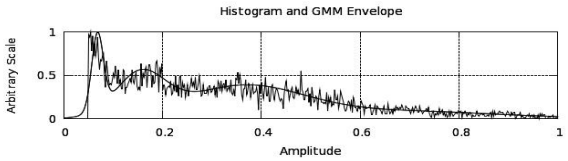

Figure 1. Comparison of the PDF envelope mixture and the normalized histogram for one example of speech signals corresponding to the word "water" spoken by a male person from the *Kamayura* community.

squared error (MSE) was used. The MSE value for the example signal is 3.004×10^{-7}. This low value shows that the GMM estimates are very similar to the expected histograms.

Fig 2 shows all the GMM PDF estimates for each of the ten signals. To construct this figure, the mixture was accomplished with four Gaussian components. It illustrates some differences among the estimated PDF, which are sustained after calculating the KL divergence and forming the dendogram.

The dendogram for all signals is shown in Fig 3 representing all the signals and their similarities or divergences. Each signal is identified using a number in the dendogram, which is the same used in Fig 2.

For the symmetrization of the KL matrix, we adopted the average means, which is performed as in (5). As previously mentioned, the consideration $\eta = 1/2$ was used, but how can it influence the estimation of the dendogram? If the values of the KL matrix are almost symmetric, the variation of η will result in little change to the dendogram, as in the case of this study. However, in other scenarios, the value of η should be chosen carefully.

V. DISCUSSION AND CONCLUSIONS

The comparison of several speech signals is not an easy task, as each signal describes a complex phenomenon that involves many parameters. This study presented an attempt to represent a hierarchical structure that shows the similarities among several indigenous words. A comparison was performed by using the KL divergence between the PDF estimates of the speech signals utilizing Gaussian mixtures and exhibiting the results with a dendogram. The Gaussian mixtures were formed using only three or four Gaussian components, which shows a satisfactory precision when compared to normalized histograms by means of MSE.

As the original signals were normalized between -1 and 1 with most values concentrated around zero, the PDF estimates using these original signals were very similar among themselves. In order to reach a more accurate and show the individual characteristics of each PDF, the estimate models were performed using the signal's envelope. The use of mixture envelope to approximate the KL divergence is an efficient algorithm, satisfying the

[3] http://www.museudoindio.org.br/

three properties of this divergence and accurately computing the KL divergence between two GMMs.

The dendogram of Fig 3 shows three main groups. If we compare these groups with the Gaussian mixtures shows in Fig 2, some similarities can be detected among

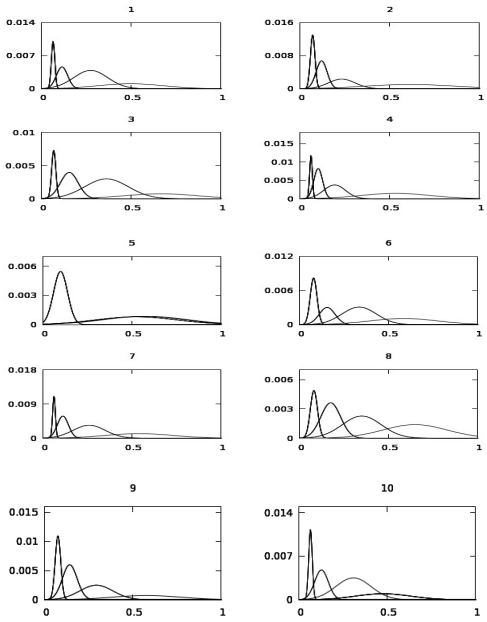

Figure 2. PDF estimates using GMM of all the speech signals. The vertical axis of each plot is on an arbitrary scale and the horizontal axis is the signal´s normalized amplitude.

the members of each group. Initially, the first group of the dendogram, with members 1, 9, and 7, the corresponding mixtures of Fig 2, present Gaussian components with lower variance values and similar mean values. The second group, with members 2, 4 and 10, has mixtures with lower variance and mean values of each Gaussian component than those of the first group. In the third group, comprising 3, 6 and 8, the mixtures are more scattered, showing higher variance values for each Gaussian component. As an outlier, the PDF estimate number 5 is different from the rest, which is clearly visible in the dendogram and also in the Gaussian mixture.

The word "water" identified in the dendogram of Fig 3 is similar for each community group. The indigenous communities *Karaja*, *Umutina* and *Kuikuro* are in the first group. The members of the second group are *Tukano*, *Karitiana* and *Yawalapiti* communities, and the last group comprises the *Kamayura*, *Kraho* and *Munduruku* communities. The word "water" in the *Kaiabi* community is different from the rest, as this community was isolated in the dendogram of Fig 3.

For future works, we intend to estimate the KL divergence of the mixtures using the parameters of each Gaussian component, as shown in [6]. Also, we intend to estimate the divergences among other words and study the resulting dendogram after variation of the η value.

This paper reports on an initial study of the subject, but as first conclusions, the technique is satisfactory in finding similarities between speech signals. Also, the KL divergence is a useful measure of difference between the models, but in the future, the aim is to use other measurements, such as mutual information.

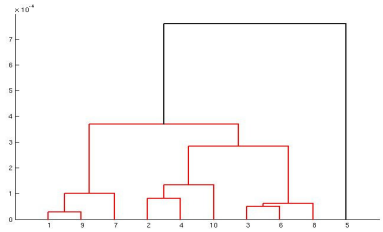

Figure 3. Dendogram that shows the similarities among the ten speech signals. This dendogram is estimated using Equation 6.

ACKNOWLEDGMENT

The authors would like to acknowledge the Brazilian Government Agency (CAPES) and the Research Foundation of SP State (FAPESP) for the financial support and scholarship given. They are also indebted to the Brazilian Indian Museum for the use of its speech database.

REFERENCES

[1] S. Gazor; W. Zhang, "Speech probability distribution," IEEE Signal Processing Letters, vol. 10, pp. 204-207, 2003.

[2] D. Erdogmus and J. C. Principe, "From linear adaptive filtering to nonlinear information processing - The design and analysis of information processing systems," IEEE Signal Processing Magazine, vol. 23, num. 6, pp. 14-33, 2006.

[3] T. Lan, D. Erdogmus, U. Ozertem and Yonghong Huang, "Estimating Mutual Information Using Gaussian Mixture Model for Feature Ranking and Selection," International Joint Conference on Neural Networks (IJCNN'06), pp.5034-5039, 2006.

[4] M.F. Huber, T. Bailey, H. Durrant-Whyte and U.D. Hanebeck, "On entropy approximation for Gaussian mixture random vectors," IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI'2008), pp. 181-188, 2008.

[5] T. M. Cover and J. A. Thomas, "Elements of Information Theory", John Wiley and Sons, Inc. New York, 1991.

[6] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2007), vol. 4, pp. 317-320, 2007.

[7] A. K. Seghouane and S. I. Amari, "The AIC criterion and symmetrizing the Kullback-Leibler divergence," IEEE Transactions on Neural Networks, vol.18, num. 1, pp. 97-106, 2007.

[8] D. H. Johnson and S. Sinanovi, "Symmetrizing the Kullback-Leibler Distance," IEEE Transactions on Information Theory, 2001.

[9] C. Tomasi, "Estimating Gaussian Mixture Densities with EM - A Tutorial," Duke University, 2004.

[10] S. Nasser, R. Alkhaldi and G. Vert, "A Modified Fuzzy K-means Clustering using Expectation Maximization," IEEE International Conference on Fuzzy Systems, pp. 231-235, 2006.

[11] G. McLachlan and T. Krishnan, "The EM algorithm and Extensions," Wiley-Interscience, 1997.

[12] G. McLachlan and D. Peel, "Finite Mixture Models," Wiley-Interscience, 2000.

[13] S. Borman, "The Expectation Maximization Algorithm - A short tutorial," University of Notre Dame, USA, 2004.

[14] J. Goldberger, S. Gordon and H. Greenspan, "An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures," Ninth IEEE International Conference on Computer Vision, vol. 1, pp. 487-493, 2003.