

GMM-Based Identification of Indonesian Speech

Amalia Zahra¹, Julie Carson-Berndsen²
*School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland*

¹Amalia.Zahra@ucdconnect.ie

²Julie.Berndsen@ucd.ie

Abstract — This paper reports the performance of identification of Indonesian speech within a ten-language corpus: English, German, Hungarian, Indonesian, Italian, Korean, Mandarin, Polish, Portuguese, and Swedish. The tasks are performed by implementing Gaussian Mixture Model (GMM) on Mel-Frequency Cepstral Coefficients (MFCCs). Two types of experiments that have been undertaken: pair-wise and ten-language experiments. In the pair-wise experiments, the performance of the model in identifying Indonesian within every language pair is evaluated. The experiments show that Indonesian is best distinguished from English, Korean, and Portuguese with 90.5% of accuracy. In ten-language experiments, the highest accuracy of identifying Indonesian is 85.71%.

Keywords — *language identification; speech processing; Indonesian identification*

I. INTRODUCTION

Language identification is a task commonly performed on either text or speech. The aim is to identify the language (but not the actual words) of the text or speech. This paper focuses on language identification on speech, and on identifying Indonesian speech in particular. However, the technique transfers to the identification of other languages.

The motivation behind the development of spoken language identification is to complement a multilingual speech recognition system, which is the ultimate future goal of the work presented here. There are two approaches in the construction of a multilingual speech recognition system [1]. The first approach is to build a single multilingual speech recognition system for all of the languages to be recognized. In this approach, language identification is implicit since once the utterance has been recognized, the language is clear. The second approach involves an explicit language identification step. The language of the speech utterance is identified first, and then the speech recognition system of that language is activated. The advantage of this approach is that the performance of the speech recognition system is the same as that of a monolingual speech recognition system as long as the language identification performs well, with a minimal amount of error. Therefore, language identification is highly important in order to build a multilingual speech recognition system with explicit language identification component.

A number of studies have been undertaken in the area of language identification [2,3,4], which address the identification

of European, East Asian, South Asian, and Middle-East languages. However, language identification studies that include South-East Asian are limited, and none of them include Indonesian. Of course, there are several studies that implement some aspects of language identification and also include Indonesian, for instance accent identification [5]. There are also studies that have proposed the utilization of corpora from other languages in order to generate data needed to build an Indonesian speech recognition system [6,7]. However, a study that specifically addresses spoken language identification for Indonesian alone has not been developed to date. Thus, the work presented in this paper is an initial step to include Indonesian speech in language identification system.

There are two main approaches to language identification using either machine learning techniques or a phone recognizer. The latter approach relies on the availability of suitably annotated corpora i.e. with word transcriptions, phonetic transcriptions, lexicon, and timing information. For some languages, such as English, German, or French, it is easier to obtain such corpora that support research and development. However, Indonesian speech corpora with complete information are more difficult to source. Therefore, machine learning techniques seem more relevant and useful for language identification where only the speech signal files and some limited annotation information are available. The relevant machine learning techniques in this domain are Gaussian Mixture Models (GMMs) [2,8], Hidden Markov Models (HMMs) [9], Neural Networks [10], k-means clustering [10], and so on. However, the GMM approach is chosen for the work presented in this paper because it has been widely used in spoken language identification [2,3,8]. This approach requires only the speech signal to be able to build a language identification system. This approach also represents an initial step in language identification process as it is envisaged that the incorporation of phone level information will be included at a later stage.

The remainder of this paper is structured as follows. Section 2 illustrates the overview of the language identification system architecture of the experiments. Section 3 explains the main GMM algorithm which is used to perform language identification. Section 4 describes the silence removal algorithm that has been utilized as a pre-processing step before the speech files are ready for use in the training and testing process of language identification. Section 5 describes gender

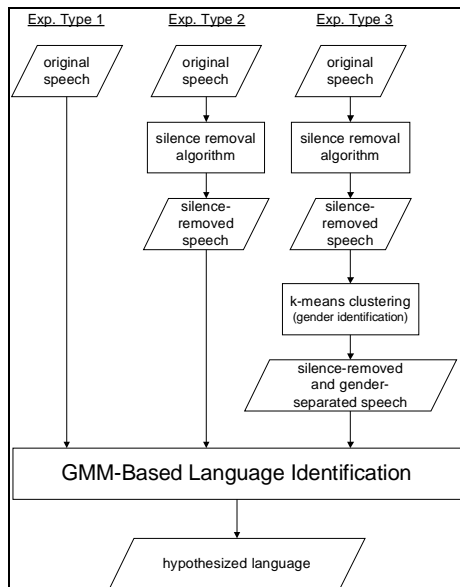


Figure 1. Language identification system architecture

classification with k-means clustering [10]. This step supports the language identification experiments on gender-separated speech. Section 6 describes the experiments carried out with respect to the identification of Indonesian speech. This section also reports the performance of each experiment with its analysis. Finally, section 7 presents several conclusions and outlines some potential future work.

II. SYSTEM ARCHITECTURE

The system architecture of the language identification process is depicted in Fig. 1. There are three types of experiments based on the types of speech files used: original speech, silence-removed speech, and silence-removed and gender-separated speech. This architecture is applied for both pair-wise (Indonesian with each of all other languages within the ten-language corpus) and ten-language experiments. The GMM method, the silence removal algorithm, and k-means clustering for gender identification will be explained further in sections 3, 4, and 5, respectively.

III. GMM-BASED LANGUAGE IDENTIFICATION

The Gaussian Mixture Model (GMM) approach has been applied in several language identification systems [2,3,8]. It takes each feature vector \vec{v}_t (at frame time t), consisting of 12 coefficients of MFCC, to build a model of a weighted sum of multivariate Gaussian densities based on (1).

$$p(\vec{v}_t|\lambda) = \sum_{k=1}^N p_k b_k(\vec{v}_t), \quad (1)$$

where λ is the set of model parameters as shown in (2).

$$\lambda = \{p_k, \bar{\mu}_k, s_k\}, \quad (2)$$

where k is the mixture index ($1 \leq k \leq N$), p_k 's are the mixture weights with the constraint that $\sum_{k=1}^N p_k = 1$, and b_k 's are the multivariate Gaussian densities defined by the means $\bar{\mu}_k$ and variances s_k .

In the identification process, an unknown speech utterance in the form of digitized speech signal is converted into MFCC feature vectors. The log likelihood of the speech utterance is then calculated for each language l model. The log likelihood L is defined in (3).

$$L(\{\vec{x}_t\}|\lambda_l) = \sum_{t=1}^T \log p(\vec{x}_t|\lambda_l), \quad (3)$$

where λ_l is the GMM for language l , T is the duration of the utterance, and \vec{x}_t are the observations. Finally the maximum-likelihood classifier hypothesizes \hat{l} as the language of the speech utterance. The hypothesized language \hat{l} is defined in (4).

$$\hat{l} = \arg \max_l L(\{\vec{x}_t\}|\lambda_l). \quad (4)$$

The GMM approach underlies the experiments described in Section 6.

IV. SILENCE REMOVAL ALGORITHM

In spoken language identification, silence is not necessarily useful since it does not contain language-specific information. Removing silence is beneficial in term of reducing the search space and the computation time. Several silence removal algorithms exist, such as Short Term Energy (STE) [11,12], Zero Crossing Rate (ZCR) [12,13], Probability Density Function (PDF) complemented with Linear Pattern Classifier (LPC) [14], signal-to-noise ratio (SNR) [2,11], and so on. STE, ZCR, and PDF complemented with LPC have been commonly applied in speaker recognition and digital signal processing, such as silence/unvoiced/voiced classification of speech and endpoint detection. In the approach presented in this paper, silence removal is utilized in a different domain, namely language identification.

The algorithm implemented here combines PDF and LPC [14] with STE [11,12]. Such a combination has been implemented in order to yield more accurate silence removal step on the corpus. The concept of PDF and LPC algorithm is to model the first 1600 samples of speech as the silence model. Then, other samples will be categorized as silence or speech based on the model. However, the silence part that usually exists at the beginning of speech file is not always long enough; sometimes the file does not even start with silence. Therefore, STE is applied within the algorithm to find the 1600 samples with the least energy as the initial silence part, with the assumption that least energy equals to silence.

The mean (μ) and standard deviation (σ) of the 1600 samples are then calculated to model their PDF. For each sample (x) of the speech files, if the one-dimensional Mahalanobis distance function¹ (z-score) is greater than 3, then the sample is categorized as speech; otherwise, it is silence. Speech is marked as 1 and silence is marked as 0. If the number of ones (speech) is greater than that of zeros (silence) in a 10-ms window, then all zeros in the window are converted into ones and vice versa. Finally, only the speech parts are selected and retrieved from the original signals in order to create a new file.

¹ Mahalanobis distance function = $\frac{|x-\mu|}{\sigma}$

The advantage of the PDF and LPC algorithms [14] is that it is not necessary to define a value as the threshold as in STE and ZCR. Although STE is applied within these algorithms, it is utilized to obtain the first set of samples to build the initial model. This approach uniquely defines the model built from the initial samples as the threshold.

V. GENDER IDENTIFICATION USING K-MEANS CLUSTERING

Gender identification has been commonly applied in speaker identification [15,16] and speech recognition [17] by using different methods. In the language identification experiments presented in this paper, gender identification is also applied in order to see whether there is an effect on the language identification performance if the training and testing data are distinguished based on the gender of the speaker.

Due to the fact that no gender information is available in the corpus, an unsupervised machine learning algorithm, one-dimensional k-means clustering [10], with k equals to 2 (male and female), is used to distinguish gender based on the pitch information of the speech.

First of all, two ($k = 2$) pitch values are selected randomly as the initial centroids: male and female centroids. Each of other pitch values is classified as a member of the male or the female cluster based on the distance between the value and each of the centroids (μ). The new mean μ_c of each cluster C_c is calculated by (5).

$$\mu_c = \frac{\sum_{x_i \in C_c} x_i}{|C_c|}, \quad (5)$$

where x_i is the pitch value of the i^{th} sample. The above steps are done iteratively until no further cluster reassignment of each value occurs.

Once the training process of k-means clustering is complete, two values are produced, each of which represents the mean pitch of each cluster. Assuming that female pitch is higher than the male pitch, the bigger value is taken to be the mean of female pitch and the smaller to be that of male pitch. These values are useful in determining which model that will be used in the language identification task, i.e. either a male or a female model.

VI. EXPERIMENTS

Experiments were carried out on the Oregon Graduate Institute (OGI) corpus [18]. Ten languages from the corpus were chosen for the language identification experiments: Indonesian, English, German, Hungarian, Italian, Korean, Mandarin, Polish, Portuguese, and Swedish. These experiments aimed to investigate the performance of GMM-based language identification at the task of identifying Indonesian from another language (pair-wise) and among the ten languages. While it would be desirable to compare the accuracy between Indonesian and other languages from the same family, namely, Austronesian, [19], such as Malay (Malaysia), Maori (New Zealand), Tagalog (Philippines), and so on, unfortunately no comparable corpora suitable for this experiment are currently readily available.

Table 1 shows the durations of the speech of each language used in the training process.

TABLE 1. TRAINING DATA

No	Language	Duration (minutes)
1	English	49.69
2	German	52.4
3	Hungarian	63.1
4	Indonesian	51.7
5	Italian	53.26
6	Korean	60.25
7	Mandarin	59
8	Polish	59.8
9	Portuguese	57.47
10	Swedish	63.24

Since the task is to see how well Indonesian speech can be distinguished from another language, just Indonesian is included in the testing process. The Indonesian testing set is different from that used in the training process. The set is divided into two types: 10-second and 45-second data set, in order to compare the language identification accuracy for short and longer speech. The numbers of 10-second and 45-second files used are 94 and 21 files, respectively.

Two types of experiments are carried out. The first experiment is pair-wise language identification, between Indonesian and each of other nine languages. The second experiment trains all ten-language speech files and then identifies Indonesian by using the model generated from the training process. Each type of experiment is implemented on 1) the original speech, 2) silence-removed speech, and 3) silence-removed and gender-separated speech.

A. Pair-Wise Language Identification

Nine experiments are undertaken regarding pair-wise language identifications. Fig. 2 shows the performance of identifying Indonesian in the pair-wise language identification for each type of speech data. It presents the accuracies of all experiments carried out on 10-second and 45-second intervals of original speech, silence-removed speech, and silence-removed and gender-separated speech.

Experiments on the original speech show that Indonesian is best distinguished from Hungarian with 84.04% and 85.71% of accuracy for 10-second and 45-second speech intervals, respectively. On 10-second silence-removed speech, the highest accuracy is obtained in Indonesian-English and Indonesian-Korean pairs with 77.7%, while on 45-second speech, it is obtained in Indonesian-English, Indonesian-Korean, and Indonesian-Portuguese pairs with 90.5%. Finally, the highest accuracy on 10-second silence-removed and gender-separated speech is 75.3%, which is produced in the Indonesian-Portuguese experiment, and that on 45-second speech is 80%, which is produced by the Indonesian-English, Indonesian-Italian, Indonesian-Portuguese, and Indonesian-Swedish experiments.

Based on the illustration in Fig. 2, the overall performance of Indonesian identification on original speech is worst. Even some of the identifications on 45-second speech are worse than those on 10-second speech, which is surprising. Typically, the longer the speech duration, the better the

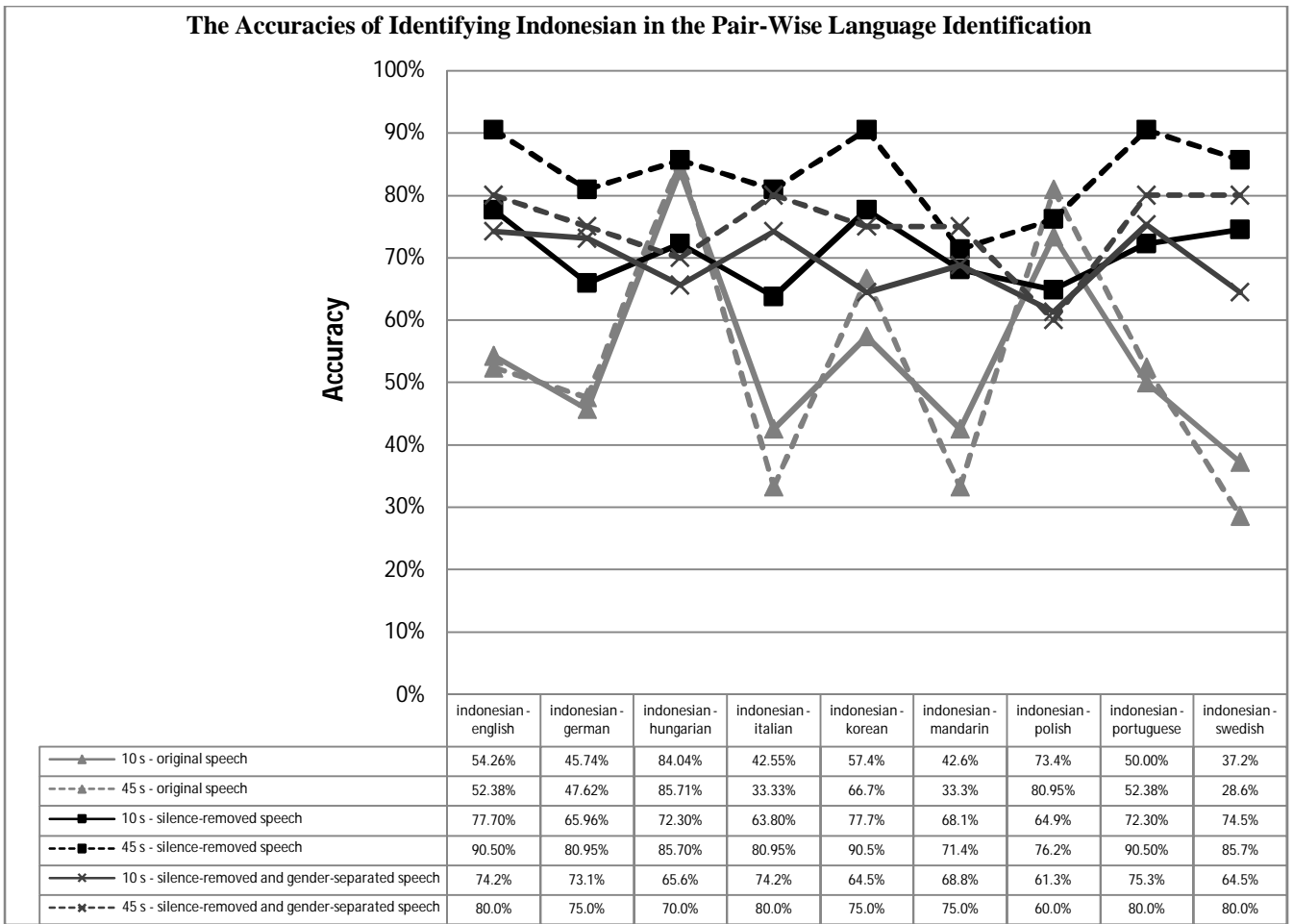


Figure 2. Indonesian Identification Accuracies in the Pair-Wise Language Identification

performance. This kind of inconsistency indicates that the silence parts of the speech file may degrade the language identification performance since they contain no language-specific information. Moreover, the accuracies of Indonesian identification using the filtered speech, either silence-removed or silence-removed and gender-separated speech, are not significantly different. The reason for this may be that the numbers of samples used in the training process for silence-removed and gender-separated speech are smaller than those used in the training process for silence-removed speech since the GMM training is separated between male and female. However, it shows consistently that almost all 45-second (filtered) speech accuracies are higher than the 10-second speech ones. From Fig. 2 it can also be seen that Indonesian is best distinguished from English, Korean, and Portuguese with 90.5% of accuracy.

B. Indonesian Identification within Ten-Language Corpus

In line with the pair-wise experiments, the ten-language experiments are also carried out on original speech, silence-removed speech, and silence-removed and gender-separated speech. Fig. 3 illustrates the accuracies of identifying Indonesian within a ten-language corpus. On 10-second speech, the figure shows 27.66%, 60.64%, and 53.76% of

accuracy for original speech, silence-removed speech, and silence-removed and gender-separated speech, respectively. While on 45-second speech, it shows 23.81%, 85.71%, and 85% of accuracy for the same speech data set.

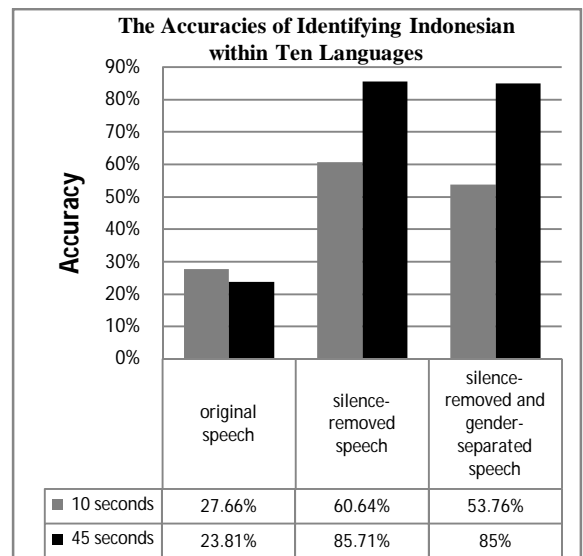


Figure 3. Indonesian Identification Accuracies within Ten Languages

Moreover, Fig. 3 shows that language identification using original speech yields the worst performance. It demonstrates that silence removal is important for language identification. The Indonesian identification accuracies are again much higher on the filtered speech (silence-removed speech and silence-removed and gender-separated speech) data although there are no significant differences between these accuracies. The highest accuracy that could be obtained in these experiments is 85.71% for 45-second silence-removed speech.

VII. CONCLUSIONS AND FUTURE WORK

This paper has presented research on language identification for Indonesian speech using a GMM on original and filtered speech. The performance of the model at the task of identifying Indonesian speech was evaluated in pair-wise and ten-language experiments. Five specific conclusions can be drawn from the experiments:

1. A Gaussian Mixture Model (GMM) may be used to perform language identification if the corpus contains nothing other than the speech itself, i.e. no word transcriptions, phonetic transcription, or lexicon.
2. The silence removal is not only effective to remove the unwanted information, but also to improve the performance.
3. In the pair-wise language identification experiments, Indonesian is distinguished the best from English, Korean, and Portuguese with 90.5% of accuracy. The analysis of why Indonesian is best distinguished from English, Korean, and Portuguese is the subject of future work.
4. The highest accuracy of Indonesian identification within the ten-language corpus is 85.71%, on 45-second silence-removed speech. The analysis behind this performance is also the subject of future work.
5. In both the pair-wise and the ten-language experiments, the accuracies of Indonesian identification are similar across the experiments on the silence-removed speech and those on the silence-removed and gender-separated speech. However, the latter experiments appear promising since the training data are smaller.

Future work will involve noise reduction on the speech data in order that the system can be used in live mode. Since noise contains no language-specific information, so it may degrade the accuracy of the system. Furthermore, based on the insights of [2,4], phonetic information will be taken into account in order to build a more accurate language identification system.

ACKNOWLEDGEMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next

Generation Localisation (www.cngli.ie) at University College Dublin (UCD). The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

REFERENCES

- [1] U. Uebler, "Multilingual speech recognition in seven languages", *Speech Communication*, vol. 35, pp. 53-69, August 2001.
- [2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no.1, January 1996.
- [3] M. A. Zissman and K. M. Berkling, "Automatic language identification", *Speech Communication*, vol. 35, pp. 115-124, August 2001.
- [4] M. Adda-Decker, et al, "Phonetic knowledge, phonotactics and perceptual validation for automatic language identification", proceeding of 15th ICPhS, Barcelona, 2003.
- [5] G. Choueiter, et al, "An empirical study of automatic accent classification", proceeding of ICASSP, Nevada, 2008.
- [6] E. Wong, et al, "Multilingual phone clustering for recognition, of spontaneous Indonesian speech utilising pronunciation modelling techniques", proceeding of 8th Eurospeech, Geneva, September 2003.
- [7] T. Martin, et al, "Cross-lingual pronunciation modelling for Indonesian speech recognition", proceeding of 8th Eurospeech, Geneva, September 2003.
- [8] P. A. Torres-Carrasquillo, et al, "Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features", proceeding of ICSLP, Denver, pp. 89-92, September 2002.
- [9] D. Jurafsky and J. H. Martin, *Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition*. xxvi, 934p, New Jersey: Prentice Hall, 2000.
- [10] T. M. Mitchell, *Machine Learning, international editions*, xvii, 414p, London: McGraw-Hill, 1997.
- [11] Q. Li, et al, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no.3, March 2002.
- [12] M. Greenwood and A. Kinghorn, "SUVing: automatic silence/unvoiced/voiced classification of speech", Department of Computer Science, University of Sheffield, UK, 1999.
- [13] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell Syst. Tech. J.*, vol. 54, pp. 297-315, February 1975.
- [14] G. Saha, et al, "A new silence removal and endpoint detection algorithm for speech and speaker recognition applications", proceeding of NCC 2005, January 2005.
- [15] G. Chetty and M. Wagner, "Multimodal speaker verification using ancillary known speaker characteristics such as gender or age", proceeding of Interspeech, Brighton, 2009.
- [16] Wang Z, et al, "Speaker gender identification based on audio fractal dimension and pitch feature", *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*, August 2008.
- [17] H. D. Zhang and J. W. Li, "Speech recognition based on CHMM classified by gender identification", *Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications)*, vol. 42, no. 21, pp. 187-189, 21 July 2007.
- [18] The OGI Multi-language Telephone Speech Corpus, Linguistic Data Consortium, 1994, <http://cslu.cse.ogi.edu/corpora/22lang/>
- [19] M. P. Lewis (ed.), "Ethnologue: languages of the world. sixteenth edition, Dallas: SIL International, 2009.