Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v35i4.19825

A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in Portuguese and English

Diego Furtado Silva*, Vinícius Mourão Alves de Souza and Gustavo Enrique de Almeida Prado Alves Batista

*Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Av. Trabalhador são-carlense, 400, 13566-590, São Carlos, São Paulo, Brazil. *Author for correspondence. E-mail: diegofsilva@icmc.usp.br*

ABSTRACT. Recognition of isolated spoken digits is the core procedure for a large number of applications which rely solely on speech for data exchange, as in telephone-based services, such as dialing, airline reservation, bank transaction and price quotation. Spoken digit recognition is generally a challenging task since the signals last for a short period of time and often some digits are acoustically very similar to other digits. The objective of this paper is to investigate the use of machine learning algorithms for spoken digit recognition and disclose the free availability of a database with digits pronounced in English and Portuguese to the scientific community. Since machine learning algorithms are fully dependent on predictive attributes to build precise classifiers, we believe that the most important task for successfully recognizing spoken digits is feature extraction. In this work, we show that Line Spectral Frequencies (LSF) provide a set of highly predictive coefficients. We evaluated our classifiers in different settings by altering the sampling rate to simulate low quality channels and varying the number of coefficients.

Keywords: spoken digit recognition, mel-frequency cepstrum coefficients, line spectral frequencies.

Um estudo comparativo entre os coeficientes MFCC e LSF no reconhecimento automático de dígitos isolados pronunciados em português e inglês

RESUMO. Reconhecimento de dígitos falados isoladamente é o procedimento fundamental para um grande número de aplicações importantes que dependem somente da fala para troca de dados, como em serviços de telefonia, tais como discagem, reserva de passagens aéreas, transações bancárias e cotações de preço. O reconhecimento é uma tarefa desafiadora visto que os sinais possuem curto período de tempo e muitas vezes alguns dígitos são acusticamente muito semelhantes a outros dígitos. O objetivo deste trabalho é investigar o uso de algoritmos de aprendizado de máquina para reconhecimento de dígitos falados e divulgar para a comunidade científica a livre disponibilidade de um banco de dados com dígitos pronunciados em inglês e português. Uma vez que algoritmos de aprendizado de máquina são totalmente dependentes de atributos preditivos para construir classificadores precisos, acreditamos que a tarefa mais importante para reconhecimento de dígitos falados é a extração de características. Neste trabalho, mostramos que *Line Spectral Frequencies* (LSF) fornecem um conjunto de coeficientes altamente preditivos. Os classificadores foram avaliados em diferentes configurações alterando a taxa de amostragem para simular canais de baixa qualidade e variando o número de coeficientes.

Palavras-chave: reconhecimento de dígitos falados, coeficientes mel-cepstrais, frequências de linhas espectrais.

Introduction

In the last decades, research on speech and speaker recognition has attracted an enormous amount of attention, mainly due to the increasing number of applications such as biometric authentication, in which a user's voice is used to allow or deny access to a system; and accessibility, in which a user is able to control equipment or navigate the Internet using speech; thus facilitating these tasks to physically impaired people.

An important speech recognition application, especially useful for telephone service providers, is is the recognition of isolated spoken digits. It can be used to replace the unattractive 'push button' system used in Interactive Voice Response menus. By using speech interaction, companies make their services user-friendlier compared with entering numbers on the telephone keypad. This is even more evident when the procedure is done through mobile devices, in which there are no physically detached keyboards for dialing.

Digit recognition seems to be an easy task compared to general speech recognition. However, spoken digit recognition is challenging due to two main reasons (KOPPARAPU; RAO, 2004):

- Spoken digits are of short acoustic duration, typically a few seconds of speech;

- Some digits are acoustically very similar to each other (for example, 'one' and 'nine').

Due to the relevance of this problem, several papers have been published, trying to improve digit recognition in different languages, such as English (NIMJE; SHANDILYA, 2011), Portuguese (SILVA et al., 2012), Arabic (ALOTAIBI, 2005) and Mandarin (SHYU et al., 2000).

Speech recognition applications, including spoken digits, follow the process of pattern recognition. In this process, summarized in Figure 1, an important step is feature extraction. This step is crucial to the application's success, since machine learning algorithms are fully dependent on predictive attributes to build precise classifiers. In speech recognition, and in sound recognition in general, the raw (audio) data are composed of a huge amount of very weak features. In this way, most machine learning algorithms are not able to build accurate classifiers, mainly due to the curse of dimensionality (FRIEDMAN, 1997). Therefore, we believe that the most important task for successfully recognizing spoken digits is feature extraction.

In general, spoken digit recognition papers have a common framework, in which Mel-Frequency Cepstral Coefficients (MFCC) are used as main features. Recently, we had discussed that Line Spectral Frequencies (LSF) coefficients provide very competitive results in the recognition of digits in Portuguese (SILVA et al., 2012).

In this paper, we expand the experiments of Silva et al. (2012) in terms of data analysis and methodology in the following aspects:

- We use a new database of spoken digits in Portuguese and English. Therefore, we evaluate the influence of language in our classification results and the potential of extrapolating our techniques to other languages;

- We provide a wider set of experimental settings with different number of MFCC and LSF coefficients. Thus, we provide a deeper understanding of the influence of such parameters in the classification performance;

- We vary the audio sampling rate to simulate the frequency response range of public switched telephone networks. Therefore, we evaluate the robustness of our method in low-quality channels such as standard telephone lines.

We note that in order to make our results fully reproducible, we made our newly collected database of spoken digits publicly available on a paper website (SILVA et al., 2013). This paper website also contains all data, code and supplemental material that were not included in this paper due to space restrictions.

Our results show that Line Spectral Frequencies (LSF) provide a set of highly predictive coefficients for digit recognition. The results are superior to those obtained with state-of-the-art methods using Mel-Frequency Cepstrum Coefficients (MFCC) for digit recognition in most settings. In particular, we show that the choice of the right feature extraction method is as important as the specific classification paradigm; and that the right combination of classifier and attributes can provide accurate classifiers for digit recognition.

Material and methods

Spoken digits database

In a previous work (SILVA et al., 2012), we proposed the use of LSF coefficients as features to classify spoken digits. In this paper we decided to build a new database of spoken digits to overcome some limitations of the previous data. More specifically, our database possesses the following features:

- 33 speakers, from 20 to 50 years old of both sexes (72.73% are male voices and 27.27% are female). The previous database had only male speakers;

- The volunteers speak the digits in a random order. In the previous database all speakers said the digits in ascending order that caused a clear intonation change in the last digit;

- The division of training and test sets is absolute. This makes the comparison with other methods simpler, and avoids the same speaker to appear in both training and test set at the same time. The previous database had no definite training and test set split;

- Digits are spoken in two languages: Portuguese and English. The previous database only includes the Portuguese language;

Figure 1. Simple pattern recognition scheme.

- Our database is segmented. Although the speaker says all 10 digits in a sequence, we split the sequence by each digit and provide a recording with each individual digit. The previous database was not segmented. As we are only interested in the performance analysis of feature extraction and classification methods, providing segmented data isolates the influence of the segmentation algorithm over the results.

Each speaker pronounced four sequences, two in Portuguese and two in English. The background noise was not controlled. Despite that all the audio files were recorded in closed rooms, such rooms are sensitive to external noise and sometimes had air conditioning or computers turned on. Some files may have rain noise or people talking in the surrounding environments. The database was separated between different languages and, for each language, divided into training and test samples. The training set consists of two thirds of the data set and the remaining one–third is used as the test set, allowing for the same experimental setup to be used in other studies. This training/test split is random; however, if one speaker is present among the test samples, he/she will not appear in the training examples or vice versa. Therefore, the classification performance is speaker independent and more consistent with real world applications.

In total, 87.88% of the volunteers are Portuguese native speakers; however, we have no English native speakers. These characteristics add challenges to the recognition task, not commonly seen in other databases. A detailed description of the set of voices is available on a website created specifically for this paper (SILVA et al., 2013).

In order to segment the database, we used a simple amplitude-based detector. This simple detector works very well because of the high signal-to-noise ratio of the data. First, we used spectrum subtraction based noise reduction (BOLL, 1979). We swept a sliding window across the signal and calculated the mean signal amplitude within each window and used that statistic as a confidence estimate. After that, the amplitude vector was normalized, dividing the values in each window by the highest value observed. Thus, the relative amplitudes will be within the intervals of 0 and 1. The confidence that the window contains part of a spoken digit is proportional to the mean amplitude. Finally, we set an acceptance threshold so that the portions above the threshold are indicative of a spoken digit. We saved the segments above the threshold in separate files. This detection method is illustrated in Figure 2.

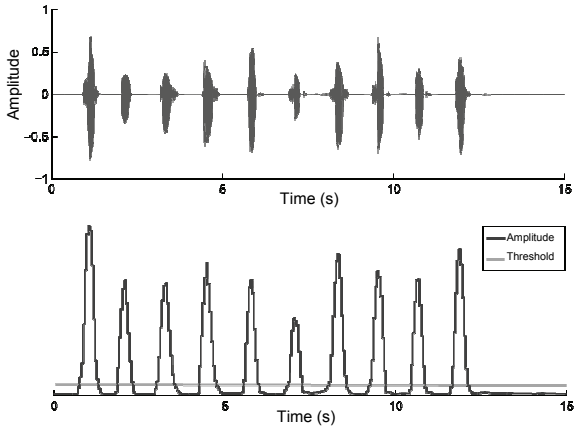

Figure 2. Segmentation scheme in which confidence values above the threshold are indicative of a spoken digit.

Feature extraction

In this section we provide a brief description of two well-known methods for feature extraction, namely Mel-Frequency Cepstrum Coefficients (MFCC) and Line Spectral Frequencies (LSF). The main purpose of these methods is to perform a representational change of the original audio data, from a high-dimensional weak-feature domain to a low-dimensional strong-feature domain.

We start describing the technique of dynamic windowing. This technique is important for two reasons: i) the window allows for extracting local features from smaller parts of the signal and characterizing signal changes in time; ii) the dynamic setting allows for adjusting the size of the window so that every signal results in the same number of features.

Dynamic windowing

Usually, speech recognition involves the classification of signals with different durations. This variability occurs not only within classes, because different spoken digits have different durations, but also between classes, because different speakers usually have different speaking paces. Data with varying length is a problem for several machine learning algorithms that expect a fixed-size attribute-value table as input. Therefore, we use dynamic windowing as a strategy to generate fixed-size attribute vectors.

Dynamic windowing is a simple strategy that breaks a signal of arbitrary length into a set of feature vectors. Each feature vector is the collection of features extracted from a segment of the original signal, which is obtained from a sliding window of width w_s. The value of w_s is dependent on the length of the signal s and the number of windows required. Furthermore, each window has an overlap with the

previous one, as illustrated in Figure 3. This overlap must be large enough so that information in the window transitions is not lost. An overlap higher than 50% is commonly used.

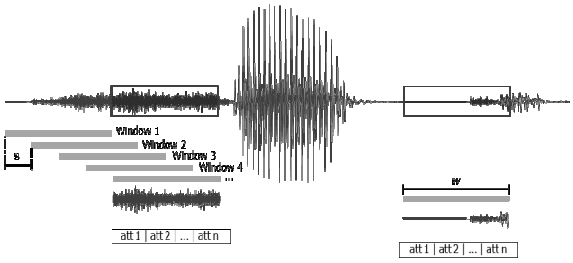

Figure 3. In the dynamic windowing strategy, the feature extraction uses a sliding window of width *w*, proportional to a preset number of windows, with an overlap of size *s* among consecutive windows. So, a *n*-dimensional feature vector is extracted for each window.

In our experiments, we set the window width w_s according to Equation 1:

$$w_s = \left\lceil \frac{e}{1-o} \right\rceil \quad (1)$$

where:

o is the overlapping rate in the interval [0, 1);

e is the window width disregarding overlapping between consecutive windows.

The value of *e* can be obtained from Equation 2.

$$e = \left\lceil \frac{l_s}{n} \right\rceil \quad (2)$$

where:

l_s is the signal length;

n is the number of windows.

The dynamic window strategy should be used with a word of caution: the existence of signals with considerable differences in duration will create a large variance in the window sizes, and consequently the step sizes. Large step sizes may cause a loss of detail on how the signal evolves with time.

Mel-Frequency Cepstrum Coefficients (MFCC)

In the past few decades, the Mel-Frequency Cepstrum Coefficients have been popularly used as features in speech processing tasks, such as speaker and speech recognition (CHIA AI et al., 2012). Briefly, to calculate those coefficients, we first take the magnitudes of frequency components using an acoustically-defined scale called 'mel'. Next, we apply a Discrete Cosine Transform (AHMED et al., 1974) on the resulting representation. The MFCC are the cepstrum coefficients obtained from this operation. Equation 3 shows the conversion from frequency (*f*) to mel-frequency (*m*).

$$m = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

Line Spectral Frequencies (LSF)

Linear Prediction (LP) is a technique used in many speech applications, such as recognition (), compression (MIELIKAINEN et al., 2010) and modeling (EDUATI et al., 2010). LP is based on the fact that a speech signal can be described by Equation 4.

$$x_k = \sum_{i=1}^{p} a_i x_{k_i} \quad (4)$$

where:

k is the time index;

p is the order of LP - ie, the number of employed LP coefficients.

The a_i coefficients are calculated in order to minimize the prediction error by means of a covariance or auto-correlation method.

Equation 4 can be rewritten in the frequency domain with a *Z*-transform (OPPENHEIM; SCHAFER, 2009). In this way, a short segment of speech is assumed to be generated as the output of an all-pole filter $H(z) = 1/A(z)$, where $A(z)$ is the inverse filter such that:

$$H(z) = \frac{1}{A(Z)} = \frac{1}{1 - \sum_{i=1}^{p}(a_i z^{-i})} \quad (5)$$

The Line Spectral Frequencies (LSF) representation, introduced by Itakura (1975), is an alternative way to represent LP coefficients. In order to calculate LSF coefficients, the inverse filter polynomial is decomposed into two polynomials: $P(z) = A(z) + z^{p+1} A(z^{-1})$ and $Q(z) = A(z) - z^{p+1} A(z^{-1})$, where $P(z)$ is a symmetric polynomial and $Q(z)$ is an antisymmetric polynomial. The roots of $P(z)$ and $Q(z)$ determine the LSF coefficients.

LSF are well suited for quantization and interpolation (KLEIJN; PALIWAL, 1995). Therefore LSF can represent the speech signal, mapping a large signal to a small number of coefficients, better than other LP representations.

Evaluation

We perform several experiments to evaluate the classification of spoken digits in Portuguese and English using MFCC and LSF coefficients. We also compare the methods on sixteen different settings,

varying the classification algorithm and their internal parameters. The algorithms we use for classification and their related settings are shown in Table 1. We chose these algorithms because they are frequently reported in the literature as having good classification performance on signal recognition tasks.

Table 1. Brief description of the classification algorithms used in our experimental evaluation.

Identifier	Inducer/setting
1-NN	1-Nearest Neighbor
3-NN	3-Nearest Neighbor weighted by inverse distance
5-NN	5-Nearest Neighbor weighted by inverse distance
7-NN	7-Nearest Neighbor weighted by inverse distance
9-NN	9-Nearest Neighbor weighted by inverse distance
11-NN	11-Nearest Neighbor weighted by inverse distance
SVM-Poly1	Support Vector Machine with Polynomial Kernel and Degree 1
SVM-Poly2	Support Vector Machine with Polynomial Kernel and Degree 2
SVM-Poly3	Support Vector Machine with Polynomial Kernel and Degree 3
SVM-RBF0.01	Support Vector Machine with RBF Kernel and Gamma=0.01
SVM-RBF0.05	Support Vector Machine with RBF Kernel and Gamma=0.05
SVM-RBF0.1	Support Vector Machine with RBF Kernel and Gamma=0.1
NB	Naïve Bayes
10 RF	Random Forest with 10 trees
15 RF	Random Forest with 15 trees
20 RF	Random Forest with 20 trees

For both approaches, MFCC and LSF, features were extracted with the previously described strategy of dynamic windowing. The width and step size of the sliding window were such that for each signal a set of *25* feature arrays was generated and each adjacent pair of windows had an overlapping area of *75%*. Therefore, each feature extraction method generated a dataset in which each instance consisted of $25 \times n$ attributes, n being the number of extracted features. For example, in a 13 MFCC scenario, each example will have 325 features.

An inherent problem in the use of MFCC and LSF coefficients as features is the choice of the number of coefficients to be used. Commonly, thirteen MFCC are used in most applications of speech recognition and speaker recognition. In order to verify whether this number provides good classification accuracy and find a number of LSF coefficients, we conducted an experiment with a large variation in the number of extracted coefficients. For each classification scenario, we perform a 10-fold cross-validation over the training dataset varying the number of coefficients from 7 to 59, in steps of 2. Once we estimated the best number of coefficients on training data, we measured the classification performance on the test set. We note that this approach of estimating parameters on training data is the correct methodology to evaluate the performance of classifiers. Papers that report the performance of classifiers selecting the best parameter values based on errors estimated on the test set have misleading, frequently overoptimistic, results (SALZBERG, 1997).

The audio was recorded with a sampling rate of 44100 Hz, also known as compact disk (CD) audio quality. However, several applications of spoken digit recognition must work over restricted sampling rates. One example is the frequency response range of public switched telephone networks, which is commonly between 300 and 3400 Hz. In order to analyze the robustness of the methods in environments where data is collected at lower frame rates, we also performed experiments with our data resampled at 20% of the original sampling rate. This means that the audio sampling rate was reduced to 8820 Hz, allowing a frequency range between 0 and 4410 Hz, similar to the ones found in telephone networks and other applications.

Results and discussion

We present in Figure 4 the classification results for all classifiers listed Table 1. These results present the accuracy obtained by each classifier after a search for the best number of MFCC/LSF coefficients using cross-validation in the training set. Summarizing the results with the number of wins and losses, MFCC obtained the overall best results for English (12 wins and 4 losses) and LSF won for Portuguese (also 12 wins and 4 losses). However, LSF obtained the highest levels of accuracy for both languages with the SVM algorithm using a polynomial kernel with a degree of 3. The best classifiers obtained 87.27% and 89.09% accuracies for English and Portuguese, respectively.

To evaluate how much the methods are susceptible to an inappropriate choice of the number of coefficients, we evaluated how accuracy varies relative to the number of coefficients. For the sake of visualization clarity, we chose to show the results for only one algorithm per learning paradigm, i.e., distance-based (k-NN), statistical (SVM), probabilistic (NB) and decision trees (RF). For each paradigm, we analyzed most accurate the classifiers' (Figure 4) behaviors. Thus, Figure 5 graphically shows the results for LSF and MFCC for the English language. We did not present the results obtained with audio in Portuguese because they were very similar to the results obtained with English.

The results show that LSF is less dependent on a particular number of coefficients than MFCC. LSF performance remains relatively constant for all classifiers in a wide range of coefficients, with a slight tendency to increase performance as the number of coefficients is increased.

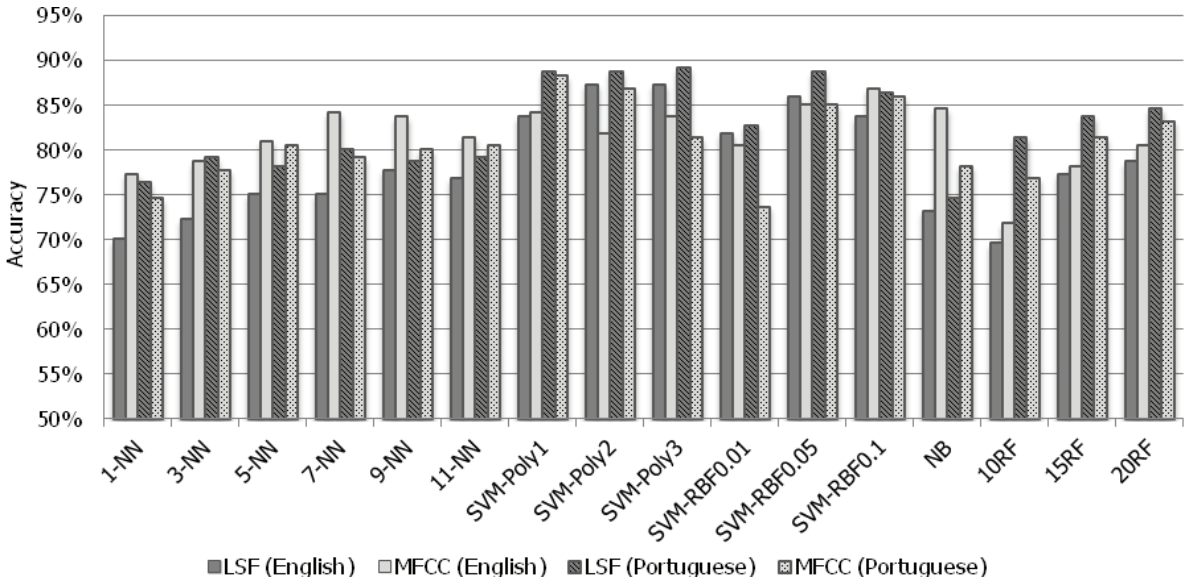

Figure 4. Spoken digit classification accuracy for English and Portuguese using LSF and MFCC.

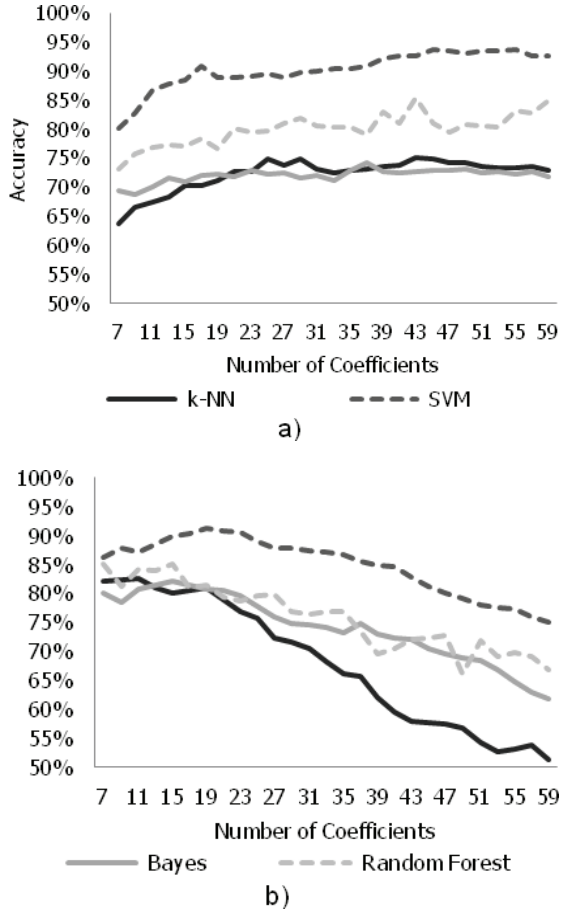

Figure 5. Variation in accuracy relative to the number of LSF coefficients (*a*) and MFCC (*b*) for English.

MFCC performance is more dependent on the correct number of coefficients, presenting an optimal value for a narrow range of coefficients and having a tendency to fall considerably as we increase the number of coefficients.

In all settings, the use of more than 25 MFCC can be considered inappropriate. In contrast, the optimal MFCC setting is in a narrow range between 11 and 23 coefficients for all classifiers, restricting the search space.

To conclude the experimental step, we conducted an experiment using the same dataset, but down sampled the data to 8820 Hz sampling rate. Again, we used cross-validation to estimate the best number of coefficients in each setting before classifying the test instances. The results are shown in Figure 6 for English and Portuguese.

It is interesting to note that in terms of accuracy, the results did not change significantly compared with the full quality signal. The best classification results for digits pronounced in both languages were achieved by SVM with a RBF kernel and 0.1 as Gamma's value. However, LSF coefficients provided the best feature space for English with 88.18% accuracy, and MFCC were the best features for Portuguese also with 88.18%.

Table 2 shows the accuracy discriminated by digit for the most accurate classifiers. We present, for each digit, its class accuracy and the most frequently misclassified digit. The results are separated by language and sampling rate.

Conclusion

In this paper, we evaluated the performance of using MFCC and LSF for the classification of digits pronounced in English and Portuguese in various scenarios.

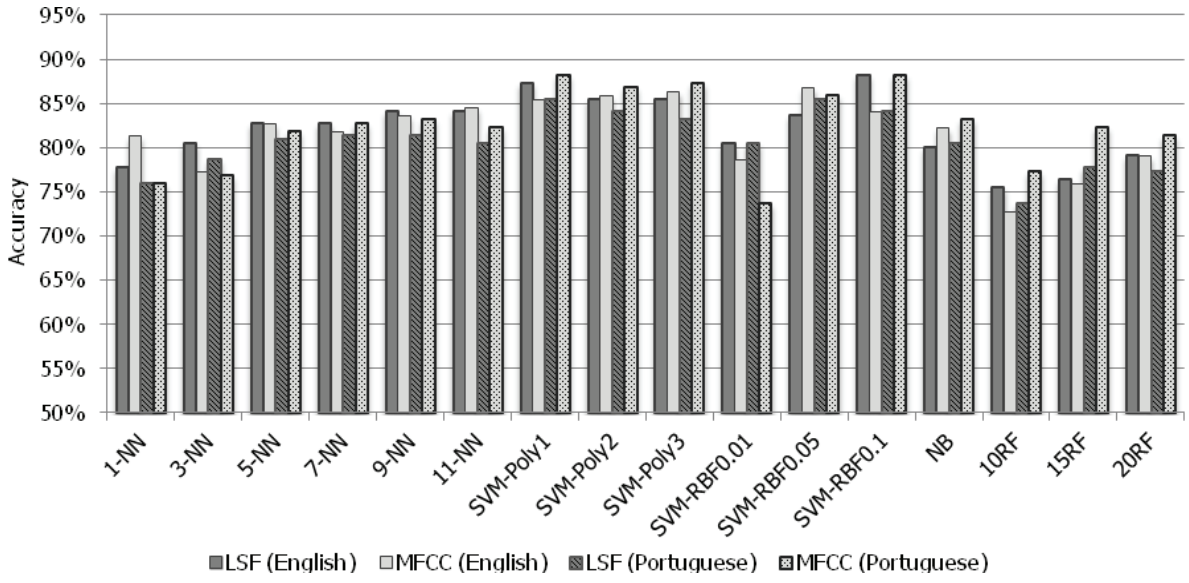

Figure 6. Spoken digit classification accuracy for English and Portuguese using LSF and MFCC with audio down sampled to 8820 Hz.

Table 2. Class accuracy and most frequently misclassified digit for the best classifiers.

	English 44100 Hz 55LSF SVM-Poly3		Portuguese 44100 Hz 41LSF SVM-Poly3		English 8820 Hz 21LSF SVM-RBF0.1		Portuguese 8820 Hz 13MFCC SVM-RBF0.1	
Digit	Accuracy (%)	Mostly misclassified as	Accuracy (%)	Mostly misclassified as	Accuracy (%)	Mostly misclassified as	Accuracy (%)	Mostly misclassified as
0	95.45	9	90.90	5 and 7	81.82	2	86.36	1, 2 and 7
1	95.45	9	86.36	8	90.90	4 and 9	100	-
2	72.73	3	90.90	3 and 9	86.36	0	77.27	8
3	90.90	2 and 8	95.45	2	77.27	8	86.36	2
4	77.27	2	86.36	9	95.45	9	90.90	9
5	81.82	9	95.45	1	81.82	9	95.45	1
6	100	-	86.36	7	81.82	3	77.27	0 and 3
7	86.36	0	95.45	0	100	-	81.82	6
8	86.36	3	86.36	1	95.45	3	100	-
9	86.36	1	77.27	4	90.90	1 and 5	86.36	1, 4 and 8

We should note that the techniques used in this paper performed similarly well for both languages. As the methods are language independent, the empirical evidence suggests that the same methodology would accurately recognize digits in other languages. However, such conjecture still requires further validation. We also presented empirical evidence that the choice of the correct number of coefficients can be as important as the correct classifier. We should note that our results only partially contribute to the general wisdom of using 13 MFCC, as it is frequently used in the literature.

Finally, the results were not significantly altered by the use of a reduced sampling rate. Therefore, we conclude that the evaluated techniques are well suited for being used in applications that require low-quality channels.

Acknowledgements

We want to thank the financial support of FAPESP (awards 2011/04054-2, 2011/17698-5, 2012/07295-3 and 2012/50714-7) and the cooperation of the volunteers who participated in the recording of the spoken digit database.

References

AHMED, N.; NATARAJAN, T.; RAO, K. R. Discrete cosine transform. **IEEE Transactions on Computers**, v. 100, n. 1, p. 90-93, 1974.

ALOTAIBI, Y. Investigating spoken Arabic digits in speech recognition setting. **Information Sciences**, v. 173, n. 1, p. 115-139, 2005.

BOLL, S. F. Supression of acoustic noise in speech using spectral subtraction. **IEEE Transactions on Acoustics, Speech and Signal Processing**, v. 27, n. 2, p. 113-120, 1979.

CHIA AI, O.; HARIHARAN, M.; YAACOB, S.; SIN CHEE, L. Classification of speech dysfluencies with MFCC and LPCC features. **Expert Systems With Applications**, v. 39, n. 2, p. 2157-2165, 2012.

EDUATI, F.; CORRADIN, A.; DI CAMILLO, B.; TOFFOLO, G. A boolean approach to linear prediction

for signaling network modeling. **PLOS ONE**, v. 5, n. 9, p. 1-6, 2010.

FRIEDMAN, J. H. On bias, variance, 0/1—loss, and the curse-of-dimensionality. **Data Mining and Knowledge Discovery**, v. 1, n. 1, p. 55-77, 1997.

ITAKURA, F. Line spectrum representation of linear predictive coefficients of speech signals. **Journal of the Acoustical Society of America**, v. 57, n. 1, p. 35-35, 1975.

KLEIJN, W. B.; PALIWAL, K. K. **Speech coding and synthesis**. Amsterdan: Elsevier Science, 1995.

KOPPARAPU, S.; RAO, P. V. S. Enhancing spoken connected-digit recognition accuracy by error correction codes. A novel scheme. **Sadhana-academy Proceedings in Engineering Sciences**, v. 29, n. 5, p. 559-571, 2004.

MIELIKAINEN, J.; HONKANEN, R.; HUANG, B.; TOIVANEN, P. Constant coefficients linear prediction for lossless compression of ultraspectral sounder data using a graphics processing unit. **Journal of Applied Remote Sensing**, v. 4, n. 1, p. 1-11, 2010.

NIMJE, K.; SHANDILYA, M. Automatic isolated digit recognition system: an approach using HMM. **Journal of Scientific and Industrial Research**, v. 70, n. 4, p. 270-272, 2011.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-time signal processing**. 3rd ed. Upper Saddle River: Prentice Hall Signal Processing, 2009.

SALZBERG, S. L. On comparing classifiers: pitfalls to avoid and a recommended approach. **Data Mining and Knowledge Discovery**, v. 1, n. 3, p. 317-328, 1997.

SILVA, D. F.; SOUZA, V. M. A.; BATISTA, G. E. A. P. A.; GIUSTI, R. Spoken digit recognition in Portuguese using Line Spectral Frequencies. In: PAVÓN, J.; DUQUE-MÉNDEZ, N.; FUENTES-FERNÁNDEZ, R. (Ed.). **Advances in artificial intelligence**. Cartagena de Índias: Springer, 2012. p. 241-250.

SILVA, D. F.; V. M. A.; BATISTA, G. E. A. P. A. **Paper website**. Retrieved February, 2013. Available from: <http://www.sites.labic.icmc.usp.br/ACTA2013/>. Access on: Aug. 29, 2013.

SHYU, R. C.; WANG, J. F.; LEE, J. Y. Improvement in connected Mandarin digit recognition by explicitly modeling coarticulatory information. **Journal of Information Science and Engineering**, v. 16, n. 4, p. 649-660, 2000.

Received on February 14, 2013.
Accepted on February 15, 2013.