





A Small Brazilian Portuguese Speech Corpus for Speaker Recognition Study

Um Pequeno Corpus em Português Brasileiro para Estudo de Reconhecimento de Locutor

Alberto Yoshihiro Nakano¹ , Hélio Rodrigues da Silva² , Juliano Rodrigues Dourado³ 
Felipe Walter Dafico Pfrimer⁴ 

Received: 3 May 2024;

Received in revised form: 5 June 2024;

Accepted: 19 June 2024;

Available online: 28 June 2024.

ABSTRACT

A small Brazilian speech corpus was created for educational purposes to study a state-of-the-art speaker recognition system. The system uses the Gaussian Mixture Model (GMM) as a statistical model for speakers and employs the Mel-frequency cepstral coefficients (MFCC) as acoustic features. The results using clean and noisy speech are compatible with the expected results, showing that the bigger the mismatch between training and test conditions, the worse the results. The results also improve with the increase in the utterance length. Finally, the obtained results can be used as baselines to compare with other speaker statistical models created with different acoustic features in different acoustic conditions.

keywords brazilian portuguese speech corpus, GMM, MFCC, speaker recognition

RESUMO

Um pequeno banco de dados de língua portuguesa falada foi criado para fins educacionais no estudo de sistema básico de reconhecimento de locutor. O sistema emprega o modelo de misturas gaussianas (GMM) como modelo estatístico e *Mel-frequency cepstral coefficients* (MFCC) como atributos/características acústicas para modelagem de locutores. Os resultados obtidos empregando amostras de teste limpas (sem ruído) e ruidosas de fala estão de acordo com o esperado, quanto maior o descasamento entre as condições de treinamento dos modelos de locutor e as condições de teste, pior será o resultado encontrado. O comprimento das amostras de teste auxilia na melhora do desempenho do sistema. Finalmente, os resultados obtidos podem ser usados como referência para comparação com resultados empregando outros atributos acústicos ou outros modelos estatísticos para modelagem de locutor.

palavras-chave corpus de português brasileiro, GMM, MFCC, reconhecimento de locutor

¹Prof. Dr, Dept. Electronic Engineering, UTFPR, Toledo, Paraná, Brazil, nakano@utfpr.edu.br

²Ms., Dept. Electronic Engineering, UTFPR, Toledo, Paraná, Brazil, helio.silva28@escola.pr.gov.br

³BSc., Dept. Electronic Engineering, UTFPR, Toledo, Paraná, Brazil, douradojuliano7@gmail.com

⁴Prof. Dr, Dept. Electronic Engineering, UTFPR, Toledo, Paraná, Brazil, pfrimer@utfpr.edu.br

Introduction

The term “corpus” was generally used to express a collection of text and audio data used in computer sciences and computational linguistics. One of the most famous collections is the Brown Corpus (Kučera & Francis, 1967), comprising American English text samples. Nowadays, a corpus is not restricted to computational linguistics and can indicate any collection such as Electromyography (EMG) (Diener et al., 2020), Electrocardiography (ECG) (Jyothi & Geethanjali, 2022), Electroencephalography (EEG) (Kuo & Lee-Messer, 2017), and video data (Zhang et al., 2021).

A speech corpus comprises utterances from both genders, recorded in neutral or emotional states and at different speeds. It is a fundamental part of speaker recognition research that aims to identify a subject from its unique utterances. With a corpus, a recognition system can be developed and further used in applications such as banking, forensics, security, real-time speaker recognition (Kinnunen et al., 2005), and emotional state recognition (Nassif et al., 2021). Although there is a lack of material in Brazilian Portuguese, there are initiatives to construct open Brazilian speech databases (Candido et al., 2023; Casanova et al., 2022; Leite et al., 2022; Paulino et al., 2018; Raso et al., 2012; Ynoguti & Violaro, 2008) that researchers can use.

In the speaker recognition task, the objective is to identify an unknown speaker out of N speakers. Speakers are statistically modeled using acoustic features and employed to verify which model out of N models probably generated a given test utterance. The most known feature for modeling is the Mel-frequency cepstral coefficients (MFCC) (Rabiner & Juang, 1993) and the state-of-the-art statistical model is the Gaussian Mixture Model (GMM) (Reynolds & Rose, 1995). The performance of the developed system depends on the test conditions. The greater the mismatch between training and test conditions, the worse the results. Additionally, the longer the utterance, the better the performance, but research has been conducted to improve the performance for short utterances (Liu et al., 2018).

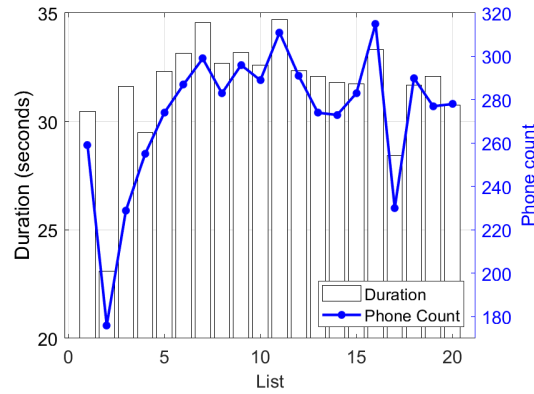
This work presents a small Brazilian Portuguese speech corpus created for didactic purposes. Experiments on speaker recognition tasks, focusing on identifying the speaker in clean and noisy conditions, are developed to define recognition performance baselines and observe degradation due to noise. Thus, section “Corpus - Data Collection and Description” describes the corpus with statistics from the database. Section “Speaker Recognition Task” presents the speaker modeling with GMM and MFCC as acoustic features employed in the experiments. The following sections present the experimental setup, results, and discussions. In the final section, we conclude and give directions on future works.

Corpus - Data Collection and Description

This study was approved by the Federal University of Technology - Paraná Ethics Committee under protocol 83708018.5.0000.5547. Forty subjects (17 men and 23 women) between 18 and 60 years were asked to naturally read and speak 20 lists of 10 phonetically balanced sentences of Brazilian Portuguese. Individual recordings were done under supervision in a quiet environment. Sentence lists were taken from (Alcain et al., 1992) which presents a study based on χ^2 distance between phone distributions that are actually observed in Brazilian Portuguese language. Recordings were done by an Olympus WS-812 digital recorder in WAVE format, at 16 bits (stereo, PCM), with a sampling frequency of 44.1 kHz. The average recording length by subject, including recording errors and long pauses, was 15 minutes, and after editing, the average length by subject was 10.5 minutes (excluding recording errors and long pauses). The total corpus length was around 7 hours.

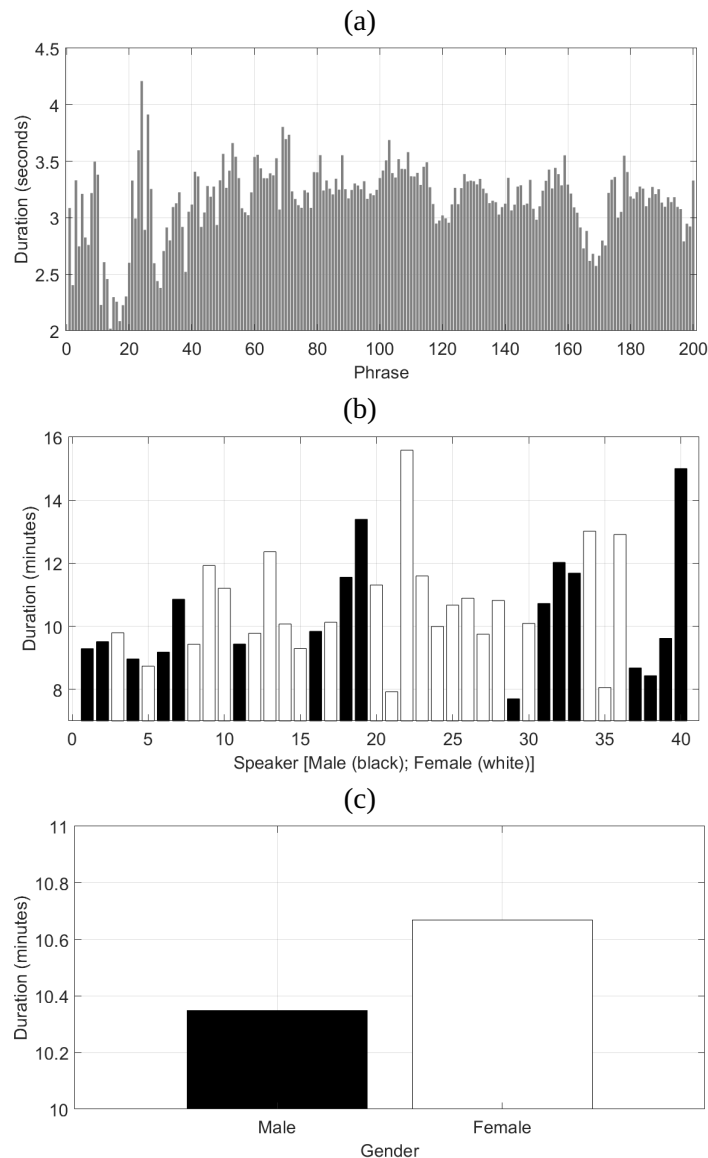
Figures 1 and 2 present some information and statistics about the corpus. In Figure 1, the correlation between the average recording time by list and the phone count by list indicates that the recording time is congruous with the list length. Figure 2 presents more information on the corpus’s average length by sentence, speaker, and gender. The average length of each spoken sentence in Figure 2(a) is consonant with the data presented in Figure 1. Figure 2(b) presents the spoken length by subject where black denotes men and white denotes female. Finally, Figure 2(c) presents the average length by gender, showing that the female group speaks slightly longer than the male group. The processed corpus on features data could be available upon contact with authors.

Figure 1 - Speech duration by list and number of phones.



From "Frequência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro" by A. Alcaim et. al, 1992, Revista da Sociedade Brasileira de Telecomunicações, 7, 40–47.

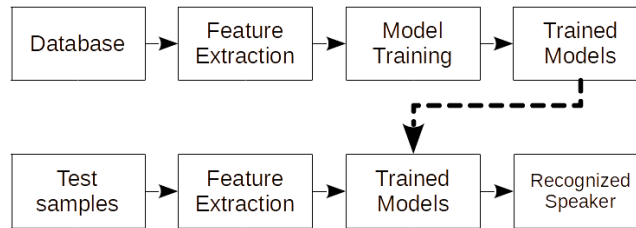
Figure 2 - Corpus statistics. (a) Average duration by phrase; (b) Average duration by speaker; and (c) Average duration by gender.



Speaker Recognition Task

The state-of-the-art speaker recognition task is based on the statistical models, the speaker model, and acoustic features. Figure 3 presents the general idea. For a given database, features are extracted and used to train speaker models, each corresponding to a different speaker. Once the models are trained, features are extracted from test samples and applied to the trained models. As a result, the recognized speaker is attributed to the speaker model, which would have generated the speech test sample. The process is discussed in the following.

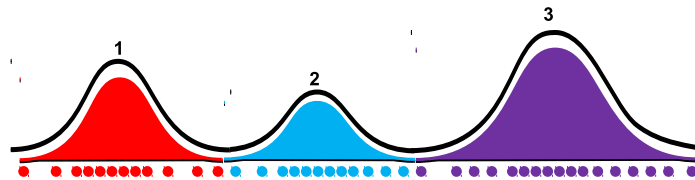
Figure 3 - Speaker recognition process.



Statistical Model - The Gaussian Mixture Model

A single Gaussian distribution can model problems described by a single variable, such as the height of people and dimensions of components manufactured by a company. However, for more complex problems, a single distribution gives a poor representation of the statistics of a data set. Figure 4 illustrates a case where a single Gaussian distribution cannot satisfactorily model the distribution. However, a good approximation can be obtained by associating three Gaussian distributions with different mean and variance.

Figure 4 - Representation of an one-dimensional distribution of a hypothetical data set.



Generalizing the idea to a D -dimensional distribution, we have

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{2\pi^{\frac{D}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)^T \right\},$$

as the probability density function where \mathbf{x} is a D -dimensional data vector, $\boldsymbol{\mu}_k$ is the mean vector with dimension D , and $\boldsymbol{\Sigma}_k$ is the covariance matrix with dimension $D \times D$. Finally, the Gaussian Mixture Model (GMM) is given by

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

which represents a weighted linear combination of k probability density functions such that weights π_k are restricted to

$$\sum_{k=1}^K \pi_k = 1,$$

and

$$\lambda = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \text{ for } k = 1, \dots, K. \quad (1)$$

In equation (1) λ represents the GMM parameters (Bishop, 2006).

Speaker Model

Let each utterance be represented by a sequence of acoustic feature vectors $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T]$, where \mathbf{x}_t denotes the acoustic feature vector of frame t . The speaker recognition task can be accomplished by computing

$$\arg \max_s \{P_r(\lambda_s | \mathbf{X})\},$$

where $P_r(\lambda_s | \mathbf{X})$ is the *a posteriori* probability of a known observation data set \mathbf{X} was generated by a speaker s . This probability is not directly computable, but using Bayes' theorem, we have

$$P_r(\lambda_s | \mathbf{X}) = \frac{p(\mathbf{X} | \lambda_s)}{p(\mathbf{X})} P_r(\lambda_s),$$

where $p(\mathbf{X} | \lambda_s)$ is the likelihood that observation \mathbf{X} came from a speaker λ_s , $p(\mathbf{X})$ is the probability density function of the observations, and $P_r(\lambda_s)$ is the *prior* probability of a speaker s modeled by λ_s in a set of speakers. In the studied case,

$$P_r(\lambda_s | \mathbf{X}) \propto p(\mathbf{X} | \lambda_s), \quad (2)$$

because observations \mathbf{X} are equally probable to be generated by any speaker, and all speakers are equally probable *a priori* to generate \mathbf{X} .

From equation (2), the speaker s recognition problem is to determine the model λ_s that probably generate observations \mathbf{X} . Lastly, the well-known Expectation-Maximization (EM) algorithm (Dempster et al., 1977) estimates parameters $\lambda_s = \{\pi_k^s, \mu_k^s, \text{and } \Sigma_k^s\}$ of each speaker s to create the speaker model $p(\mathbf{X} | \lambda_s)$. Thus, a speaker s can be associate to a speaker model represented by the GMM on equation (1).

Acoustic features

A feature tries to give a non-redundant and compact representation of information. A simple example is the analysis of one second of a pure harmonic component sampled at 16 kHz. In the time-domain, 16 thousand samples should be stored and analyzed, but in the frequency-domain all information is resumed into three features: amplitude, phase, and frequency, resulting in 99.98125 % of data compactness.

Mel-frequency cepstral coefficients (MFCC) is a popular and standard acoustic feature chosen for use in the experiments. MFCC can be obtained by applying to a signal: segmentation into frames, pre-emphasis, windowing, Discrete Fourier Transform (DFT), log-magnitude, Mel-scale filtering, Discrete Cossine Transform (DCT), and liftering. Cepstral Mean Normalization (CMN) can be applied to reduce noise effects. Additionally, the logarithm of the frame energy (logE) and first-order and second-order derivatives of the features, known as dynamic features, denoted as Δ and $\Delta\Delta$ can be included as support features.

Experiments

Preliminary experiments were conducted on the proposed speech corpus where GMM was used as the speaker model and MFCC, logE, and its dynamic features were used as acoustic features. The speech corpus was split into two datasets, 80 % in the training dataset and 20 % in the test, with no overlap between datasets. All speaker models were trained with clean speech. Acoustic feature extraction, GMM model training, and tests were performed using Matlab (Mathworks, 2024). GMM as speaker models were modeled with the following number of Gaussian components: 1, 2, 4, 8, 16, 32, 64, 128, and 256. The following features associations were tested:

- MFCC only (13 features);
- MFCC and logE (14 features);
- MFCC, logE, Δ MFCC, Δ logE (28 features); and
- MFCC, logE, Δ MFCC, Δ logE; $\Delta\Delta$ MFCC, $\Delta\Delta$ logE (42 features).

Tests were performed on clean speech and noisy speech to observe the performance of the speaker recognition system. Noisy data were artificially generated using Additive Gaussian White Noise (AWGN) on clean speech considering signal-to-noise ratio (SNR) from -20 to +40 dB with 5 dB steps. Experiments on clean speech have the objective of observing recognition rate baselines. Noisy tests are performed to verify the effect of the mismatch between training and test conditions compared to clean speech case.

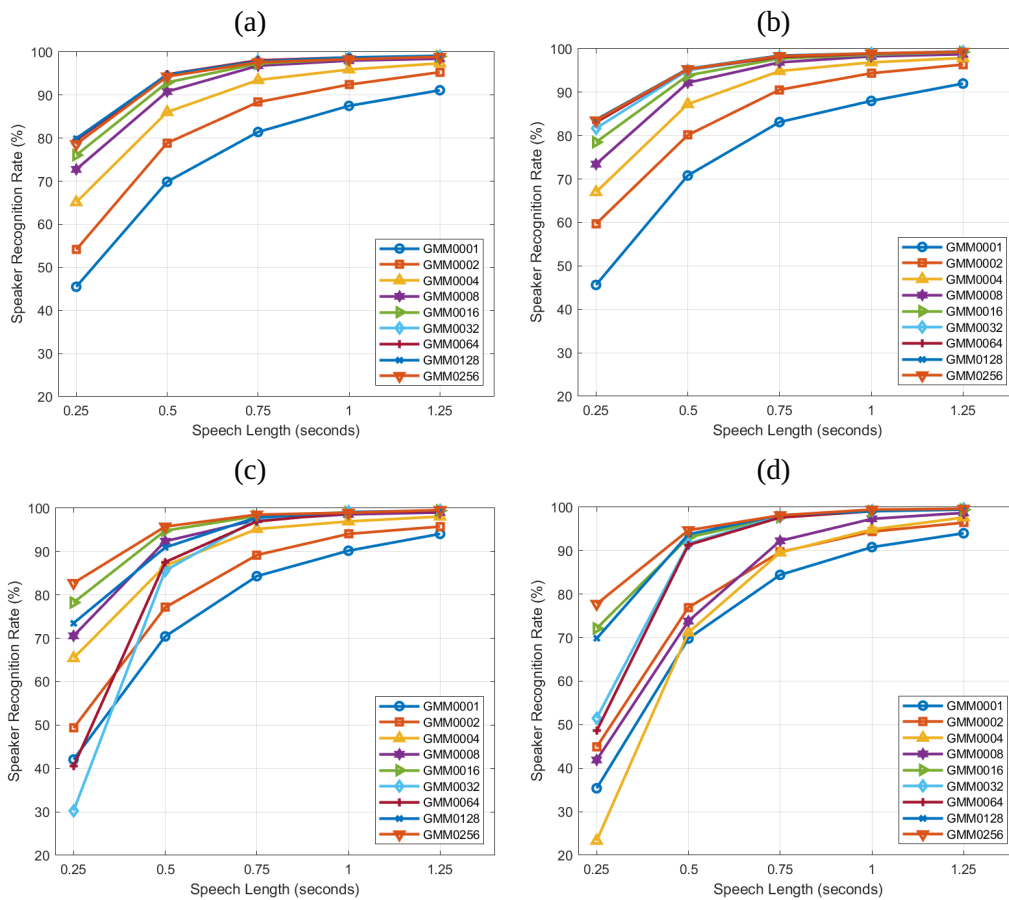
Results and Discussions

Experiments on clean speech verify the recognition rate performance by varying the spoken sentence length and the number of Gaussian components. In noisy speech, the performance is verified by increasing the spoken length at different SNR levels.

Recognition rate by spoken length on clean speech

Figures 5(a)-(d) show the recognition rate by increasing the spoken duration from 0.25 to 1.25 seconds (s) in steps of 0.25 s considering only MFCC, MFCC+logE, MFCC+logE+ Δ , and MFCC+logE+ Δ + $\Delta\Delta$ feature sets, respectively. Generally, increasing the length and the number of Gaussian components improves the recognition rate because it is more likely that distinct characteristics of the speaker are considered in the recognition process, reducing ambiguity. Additionally, the longer the sentence with voiced active sections, the shorter the silence and pause sections will be. Thus, more information is used to make a decision.

Figure 5 - Recognition by speech length on clean speech: (a) only MFCC; (b) MFCC+logE; (c) MFCC+logE+ Δ ; and (d) MFCC+logE+ Δ + $\Delta\Delta$

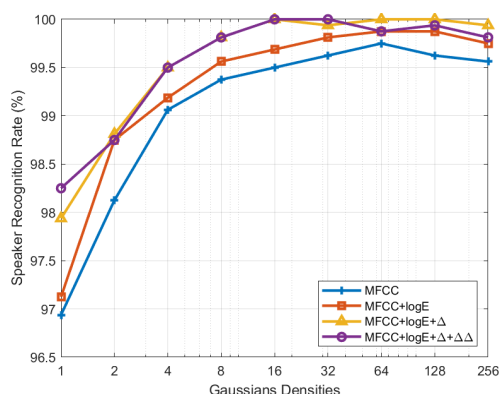


The performance of a GMM using dynamic features can be observed in Figures 5(c)-(d) by testing MFCC+logE+ Δ and MFCC+logE+ Δ + $\Delta\Delta$ sets. By adding dynamic features a more complete and complex model is generated, which is dependent on the utterance length, that is, the recognition rate is enhanced with the increase in utterance length. In the same way, by increasing Gaussian components, unpredictable results for short utterances turn out to be trustful with the length increasing. At last, the acceptable spoken length will define the trade-off between the complexity of the system and the desired recognition performance. For a low complexity system, a long spoken sentence will be needed, while for a high complex, a short sentence will be required.

Recognition by Gaussian Components on clean speech

Figure 6 presents the recognition performance using full sentence length. There is an improvement in the recognition rate by increasing Gaussian components, which is similar to what was observed in Figure 5. The contribution of dynamic features is evident when considering a low number of densities. Recognition results are expected to reach high performance as the number of densities increases because recordings were done in controlled conditions. A recognition rate higher than 97 % is obtained, compared to Figure 5, because there was no restriction on the length of the utterance. Observing MFCC and MFCC+logE curves, including logE as a feature improves the recognition rate. Including Δ and $\Delta\Delta$ shows some improvements when using a small number of densities. Apparently, including Δ and $\Delta\Delta$ gives the wrong idea that it does not provide significant gain over MFCC+logE case for higher number of densities. However, if tests are in noisy conditions, dynamic features improve results.

Figure 6 - Recognition by increasing the number of Gaussian components



Performance on Noisy Conditions

Figures 7 and 8 present the recognition performance in noisy conditions taking short test utterances of 0.25 s, and all long utterances, respectively. The performance is highly deteriorated for low SNR levels, regardless of the number of Gaussian components, the feature set, or the utterance length. Slightly improvement starts at +5 dB and increases with the SNR level. In Figures 7(a)-(d), feature sets for short utterance length, the maximum recognition rate at +40 dB is around 70 % and is not so dependent on the number of Gaussian components. Feature sets do not consistently perform well even with high SNR, which shows that noise can compromise the classification of the correct speaker in short utterances. However, increasing the utterance length results show a sigmoidal-shaped response, as seen in Figures 8(a)-(d), whose stationary phase reaches more than 99 % of recognition. This behavior matches what was discussed in the clean speech case, that is, by increasing the utterance length, it is more likely that speaker information is considered in the recognition, thus improving the results even in noisy conditions.

Noise in the test samples implies a mismatch between training and test conditions. The higher the SNR, the lower the mismatch, which means performance tends to reach the clean speech case. Decreasing the SNR level increases the mismatch, so the results worsen. It is worth noting that CMN was used in the training phase, which helped to reduce the mismatch. Without CMN recognition rate, the performance does not tend to the clean speech case even at +40 dB.

Figure 7 - Recognition on noisy conditions considering 0.25 s of speech and different features associations

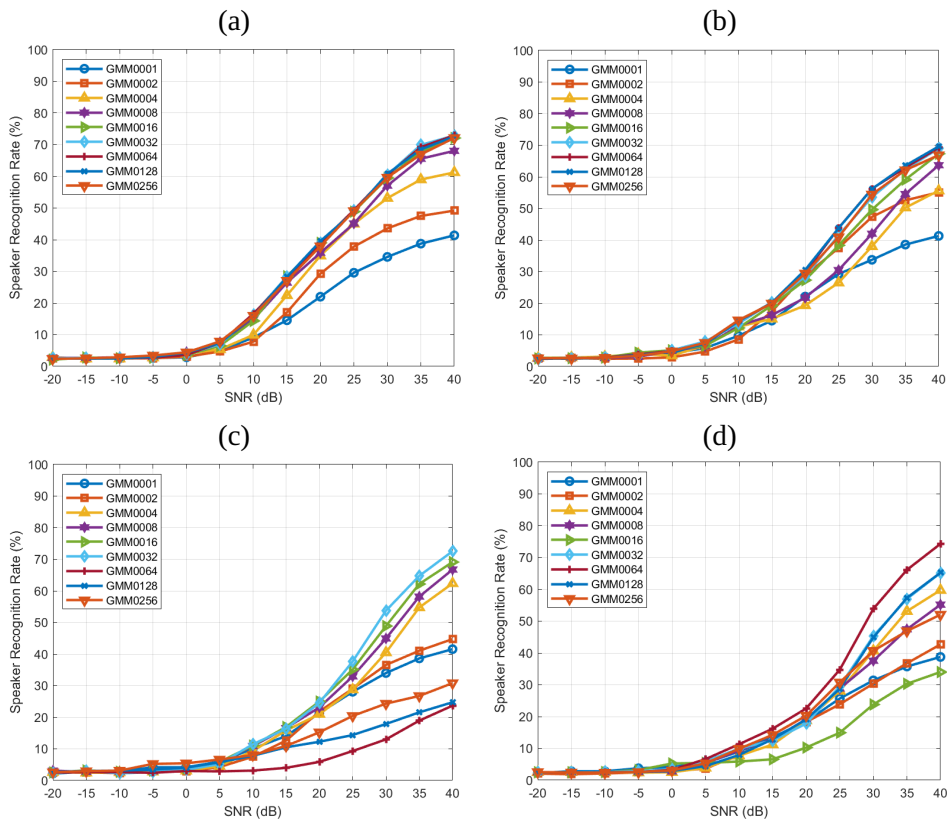
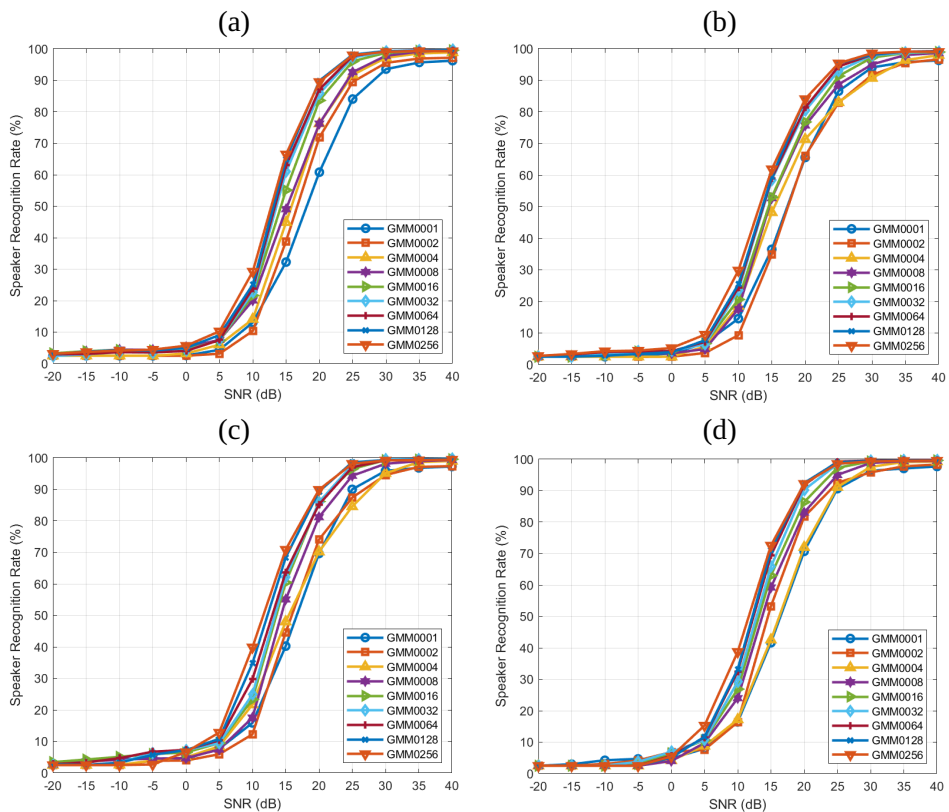


Figure 8 - Recognition on noisy conditions considering total spoken length and different features associations



Conclusions

In this work, a small Brazilian Portuguese speech corpus was created to develop a speaker recognition system based on GMM using MFCC and its dynamic features as acoustic features. Clean speech for short and long utterances was tested, establishing system baselines. Noisy test data was artificially generated to observe performance on trained GMM models with clean data. Results were compatible with the expected mismatch between training and test conditions. Finally, the created framework can be used for more research, such as evaluating under reverberant or other noisy conditions.

Author contributions

A. Y. Nakano participated in: Conceptualization, Formal Analysis, Investigation, Visualization, Supervision, Writing - original draft, revision and editing. H. R. Silva and J. R. Dourado participated in: Formal Analysis, Investigation, Methodology, Validation, Programs, Writing - original draft. F. W. D. Pfrimer participated in: Validation, Visualization, Writing - revision and editing.

Conflicts of interest

The authors certify that no commercial or associative interest represents a conflict of interest concerning the manuscript.

References

- Alcain, A., Solewicz, J. A., & Moraes, J. A. (1992). Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, 7(1), p. 40–47. <https://doi.org/10.14209/jcis.1992.2>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Candido, A. J., Casanova, E., Soares, A., Oliveira, F. S., Oliveira, L., Fernandes, R. C. J., Silva, D. P. P., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., & Aluísio, S. M. (2023). CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*, 57, 1139–1171. <https://doi.org/10.1007/s10579-022-09621-4>
- Casanova, E., Candido, A. J., Shulby, C. D., Oliveira, F. S., Teixeira, J. P., Ponti, M. A., & Aluísio, S. M. (2022). TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. *Language Resources and Evaluation*, 56, 1043–1055. <https://doi.org/10.1007/s10579-021-09570-4>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1), 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Diener, L., Vishkasougheh, M. R., & Schultz, T. (2020). CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion, In Isca Archive, *Papers [Proceedings]*. Proceedings Interspeech 2020. Shanghai, China. <https://doi.org/10.21437/Interspeech.2020-2859>
- Jyothi, S., & Geethanjali, N. (2022). Arrhythmia prediction from high dimensional electrocardiogram's data corpus using ensemble classification. *International Journal of Health Sciences*, 6(S1), 4790–4810. <https://doi.org/10.53730/ijhs.v6nS1.5898>
- Kinnunen, T., Karpov, E., & Franti, P. (2005). Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), 277–288. <https://doi.org/10.1109/TSA.2005.853206>
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press. <https://doi.org/10.1086/465045>

- Kuo, J., & Lee-Messer, C. (2017). The stanford EEG corpus: A large open dataset of electroencephalograms from children and adults to support machine learning technology. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–2. <https://doi.org/10.1109/SPMB.2017.8257034>
- Leite, P. H. L., Hoyle, E., Antelo, Á., Kruszielski, L. F., & Biscainho, L. W. P. (2022). A Corpus of Neutral Voice Speech in Brazilian Portuguese. In V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, & H. Pinto (Eds.), *Computational Processing of the Portuguese Language* (pp. 344–352). 15th International Conference, PROPOR 2022, Fortaleza, Brazil. https://doi.org/10.1007/978-3-030-98305-5_32
- Liu, Z., Wu, Z., Li, T., Li, J., & Shen, C. (2018). GMM and CNN hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial Informatics*, 14(7), 3244–3252. <https://doi.org/10.1109/TII.2018.2799928>
- Mathworks. (2024). MatLab - Designed for the way you think and the work you do. <https://www.mathworks.com/products/matlab/>
- Nassif, A. B., Shahin, I., Hamsa, S., Nemmour, N., & Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. *Applied Soft Computing*, 103, 1–11. <https://doi.org/10.1016/j.asoc.2021.107141>
- Paulino, M. A., Costa, Y. M., Britto, A. S., Svaigen, A. R., Aylon, L. B., & Oliveira, L. E. (2018). A Brazilian speech database. In *IEEE Conferences [Proceedings]. 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. Volos, Greece, 234–241. <https://doi.org/10.1109/ICTAI.2018.00044>
- Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall.
- Raso, T., Mello, H., & Mittmann, M. M. (2012). The C-ORAL-BRASIL I: Reference corpus for spoken Brazilian Portuguese. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 106–113). 8 International Conference on Language Resources and Evaluation. Istanbul, Turkey.
- Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72–83. <https://doi.org/10.1109/89.365379>
- Ynoguti, C. A., & Violaro, F. (2008). A Brazilian Portuguese speech database. In *Sociedade Brasileira de Telecomunicações, SBRT2008 [Proceedings]. XXVI Simpósio Brasileiro de Telecomunicações*, Rio de Janeiro, Brasil. <https://doi.org/10.14209/sbrt.2008.42398>
- Zhang, H., Sun, A., Jing, W., Nan, G., Zhen, L., Zhou, J. T., & Goh, R. S. M. (2021). Video Corpus Moment Retrieval with Contrastive Learning. *ArXiv*, 1, 1–11. <https://doi.org/10.1145/3404835.3462874>