LSTM Network based on Prosodic Features for the Classification of Injunction in French Oral Utterances

Asma Bougrine[1] [a], Philippe Ravier[1] [b], Abdenour Hacine-Gharbi[2] [c] and Hanane Ouachour[1]

[1] *PRISME Laboratory, University of Orleans, 12 Rue de Blois, 45067 Orleans, France*

[2] *LMSE Laboratory, University of Bordj Bou Arréridj, Elanasser, 34030 Bordj Bou Arréridj, Algeria*

Keywords: Speech Injunction Classification, Massive Wild Oral Corpus, Prosodic Features, Static and Dynamic Features, SVM, K-NN, Long Short Term Memory (LSTM).

Abstract: The classification of injunction in french oral speech is a difficult task since no standard linguistic structure is known in the french language. Thus, prosodic features of the speech could be permitted indicators for this task, especially the logarithmic energy. Our aim is to validate the predominance of the log energy prosodic feature by using conventional classifiers such as SVM or K-NN. Second, we intend to improve the classification rates by using a deep LSTM recurrent network. When applied on the RAVIOLI database, the log energy feature showed indeed the best classification rates (CR) for all classifiers with $CR = 82\%$ for SVM and $CR = 71.42\%$ for K-NN. When applying the LSTM network on our data, the CR reached a not better value of 79.49% by using the log energy feature alone. More surprisingly, the CR significantly increased to 96.15% by using the 6 prosodic features. We conclude that deep learning methods need as much data as possible for reaching high performance, even the less informative ones, especially when the dataset is small. The counterpart of deep learning methods remains the difficulty of optimal parameters tuning.

1 INTRODUCTION

The injunctive values are very useful for many studies dealing with oral speech interactions, e.g. in the field of social robotics research, automatic meaning processing or in the science of language pathology understanding and therapy. An injunction can be defined as any utterance intended to obtain from the interlocutor that he behaves according to the speaker's desire, whether it is an order or a defense. In french language, the mode of injunction is, essentially, the imperative. It can be used for any injunctive purpose from the lightest (solicitation) to more peremptory (order). However, the imperative form does not only express the injunctive value and it can be used for other purposes such as the expression of a condition or a question, etc. Then, the injunction do not have a specific structure. It can be expressed in both imperative and indicative grammatical modes. Thus, automatic classification of injunction in oral utterances is not an easy task because of the lack of an unique linguistic structure or signature. In a previous work conducted by Abdenour Hacine-Gharbi et al. (Hacine-Gharbi and Ravier, 2020), authors focused on the contribution of prosodic features to identify the injunctive values in oral speech utterances. This study has investigated some hand-engineered features such as prosodic descriptors (pitch, energy) with their associated dynamic features and other typical descriptors, such as spectral features namely Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction coefficients (PLP) and Mel-Frequency Cepstrum Coefficients (MFCC) with their first derivatives and second derivatives for dynamical modelling. Authors showed that the prosodic features are relevant for the task of classification into injunction (INJ) and no-injunction (NINJ) classes. The study also highlighted the predominance of the log energy feature. The last result is interesting and needs however to be confronted to other types of classification system for confirmation.

The confirmation of this last result is the first objective of the present paper. To validate the predominance of the logarithmic energy prosodic feature, we use other conventional classifiers such as the support vector machine (SVM) and the k-nearest neigh-

[a] https://orcid.org/0000-0002-8264-6360
[b] https://orcid.org/0000-0002-0925-6905
[c] https://orcid.org/0000-0002-7045-4759

Bougrine, A., Ravier, P., Hacine-Gharbi, A. and Ouachour, H.
LSTM Network based on Prosodic Features for the Classification of Injunction in French Oral Utterances.
DOI: 10.5220/0010910500003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 730-736
ISBN: 978-989-758-549-4; ISSN: 2184-4313

bors (K-NN) with majority-voting rule decision strategy. The second objective is to improve the classification rate by using a recurrent deep learning network namely the Long Short-Term Memory (LSTM) network. An LSTM network allows to input sequence data into a network, and make predictions based on the individual time steps of the sequence data. Thus, we pass the prosodic features to an LSTM network to attempt to improve the classification results. The remainder of the paper is organized as follows. Section 2 briefly describes the previous work presented in (Hacine-Gharbi and Ravier, 2020). Section 3 details methods used to classify the speech injunction utterances in oral speech corpus. The massive wild oral corpus RAVIOLI is described and detailed in Section 4 followed by experimental results obtained on a dataset extracted from RAVIOLI. Finally Section 5 concludes the paper.

2 RELATED WORK

The prosodic features have shown their importance in several patterns recognition systems that require a training phase for class's modeling and a testing phase for performance evaluation. Each phase requires features extraction step which divides each input signal in sequences of overlapped windows and next converts each window into a vector of features. Particularly, the energy and the pitch with their dynamics features have been used as prosodic features in automatic meaning processing (Hacine-Gharbi et al., 2015)(Hacine-Gharbi et al., 2017)(Hacine-Gharbi and Ravier, 2020). In (Hacine-Gharbi et al., 2015), the authors have proposed the use of these features for automatically classifying the speech signal of the French word 'oui' into tow classes labeled 'conviction' and 'No-conviction', using the Hidden Markov Model (HMM) classifier. The features selection based on wrappers approach has been applied on this set of features for studying their relevancy. The results have shown the relevancy of the dynamic prosodic feature noted Δf0 and the energy for this classification task. In (Hacine-Gharbi et al., 2017), the authors have studied the relevancy of local and global prosodic features using several algorithms of features selection based on the mutual information. The local features have been extracted from each phoneme of the word 'oui'. The features considered are the mean and the standard deviation of the previews prosodic features with the addition of the duration. The results have shown the superiority of the local features with respect to the global ones. In (Hacine-Gharbi and Ravier, 2020), the authors have proposed a novel framework for classifying the speech signal into injunction (INJ) and no-injunctions (NINJ) classes modeled each by GMM model (Gaussian Mixture Models) (Reynolds, 2009) during the training phase. The authors have used a subset data of 197 injunctive utterances and 198 non-injunctive utterances extracted from database of RAVIOLI project. A comparative study has been performed between several types of features used in speech recognition task (LPCC, PLP and MFCC) and the prosodic features with their dynamic features. The testing phase consists to convert each speech utterance in sequence of features vectors and next classify the whole sequence in one class using the GMM classifier which is considered as HMM with one class. The flowchart is recalled in Figure 1. The results have shown that the prosodic features are more relevant than the acoustic ones. Particularly, the application of the features selection based on the wrappers approach has demonstrated the relevancy of log energy feature. Hence, the authors have proposed the first concept of classifying the speech utterances in injunction and non-injunction classes.

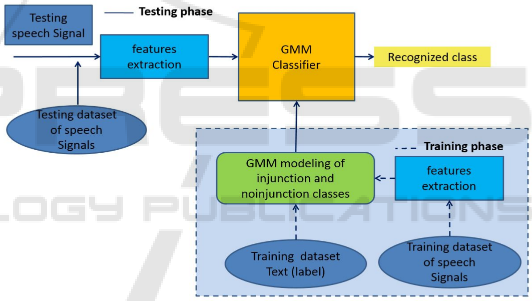
Testing speech Signal
Testing phase
features extraction
GMM Classifier
Recognized class
Testing dataset of speech Signals
Training phase
GMM modeling of injunction and noinjunction classes
features extraction
Training dataset Text (label)
Training dataset of speech Signals

Figure 1: The diagram of the classification system in the two classes INJ or NINJ using GMM classifier.

The objective of the present work consists to further demonstrate the effectiveness and the robustness of this concept using several classifiers. In this work, the same distribution of the database in training and testing sets used in the previews work is employed. Two novel classification systems are proposed for this task. The first one consists to convert each input speech utterance in sequence of prosodic features and next classify each features vector using classical classifier such as SVM and K-NN. The decision about the class of the signal is performed using the voting rule applied on the sequence of classes of features vectors. The second system proposed classifies each input speech utterance into INJ or NINJ classes using the same features with the deep learning concept.

3 THE CLASSIFICATION OF SPEECH INJUNCTION

Remember that in this study, we will consider prosodic features only as it has been approved for the task of speech injunction classification in (Hacine-Gharbi and Ravier, 2020). In this same work, the logarithmic energy feature has achieved the best classification rate. Our first aim is to evaluate the relevance of this finding. The prosodic features are the log energy 'E' values extracted from HTK software (Young et al., 1999) and the pitch 'PI' values that are computed each 10 ms by Praat software (Boersma and Weenink, 2013) on 30 ms overlapping analysing windows. The 'D' and 'A' dynamic features are calculated for these features using 'Hcopy' command of HTK Tools library. These prosodic features are then fed into two conventional learning models namely SVM and K-NN. Also, we apply a deep learning model, namely the LSTM network that takes as input a vector of prosodic features.

3.1 Conventional Learning Methods

3.1.1 Flowchart of the New Classification Systems

Two supervised classification process are performed in this work. They all require a learning phase in order to obtain a trained model for each class (INJ/NINJ) followed by a testing phase in order to identify the class of an unknown utterance signal. Thus, the dataset is divided into a training dataset and a testing dataset. Figure 2 illustrates the diagram of the automatic classification system of speech occurrences into injunctive and non-injunctive classes where the classifier bloc can be the SVM or K-NN methods. The training phase has the objective of modeling each class using an SVM or K-NN algorithm. This phase allows to create a matrix in which each line represents a prosodic features vector and the last column represents the belonging class (label) of each vector. The testing phase has the objective to evaluate the system performances using classification rate. This phase consists firstly to convert each speech utterance in sequence of features vectors and next classify each vector using one classical classifier (K-NN or SVM). The final decision about utterance class is obtained by the majority voting rule applied on the sequence of classes. The structure of this scheme has been applied in several domains such as (Ghazali et al., 2019) (Ghazali et al., 2020) (Bengacemi et al., 2021) (Ghazali et al., 2021).

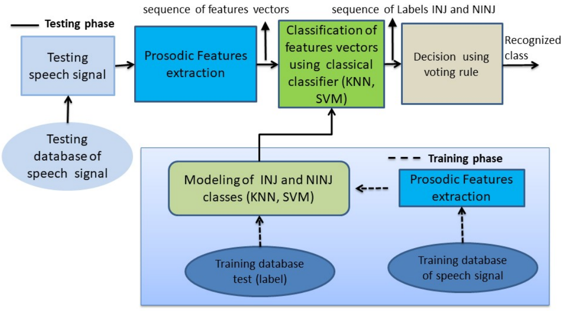
sequence of features vectors
sequence of Labels INJ and NINJ
Testing phase
Testing speech signal
Prosodic Features extraction
Classification of features vectors using classical classifier (KNN, SVM)
Decision using voting rule
Recognized class
Testing database of speech signal
Training phase
Modeling of INJ and NINJ classes (KNN, SVM)
Prosodic Features extraction
Training database test (label)
Training database of speech signal

Figure 2: The proposed system using two classifiers combined with the voting rule method.

3.1.2 Classification by SVM

SVM or Support Vector Machine (Cortes and Vapnik, 1995) is a linear and supervised classification model that can solve linear and non-linear classification problems. The idea of SVM is to find a hyperplane which separates the data into classes. If the two classes are not linearly separable, this is where the kernel approach comes in. A kernel is a function that takes the original non-linear problem and transforms it into a linear one within the higher-dimensional space. There are several kernel functions: polynomial, sigmoidal and Radial Basis Function (RBF) kernels. Here, we used a RBF kernel function. RBF is the default kernel used within the sklearn's SVM classification algorithm and can be described as: $\exp^{-\gamma||x-x'||^2}$. Two hyper-parameters have to be set γ and C in order to control the estimation error of SVM. More precisely, the hyper-parameter γ controls the tightness of the RBF function and the hyper-parameter C controls the margin width.

3.1.3 Classification by K-NN

K-Nearest neighbour (Fix and Hodges, 1989) is a non-parametric supervised machine learning method for the classification of a data sample. First, the feature vectors and class labels of the training samples are trained and stored. In the classification phase, an unlabeled vector X is classified by assigning the label which is most frequent among the k training samples nearest to X. The classes of these neighbours are weighted using the similarity between X and each of its neighbours measured by the Euclidean distance metric. The hyper-parameter k controls the extent of the neighbourhood in the decision making.

3.2 Deep Learning Methods

Because of their internal memory, recurrent neural networks (RNNs) in particular Long Short-Term

Memory Networks (LSTMs) can remember important information about the previous input they received. This is why they are frequently used for handling sequential data such as speech data.

We have used LSTM network which is an extension of RNN, introduced by Hochreiter and Schmidhuber in 1997 (Hochreiter and Schmidhuber, 1997), in order to overcome the vanishing and exploding gradients problems in RNN caused by back-propagation through time. Its cells control which information must be remembered in memory and which are not. In addition, it uses a summation of memory status instead of a multiplicative status for RNN.

Figure 3 shows the configuration of the proposed LSTM model. The size of the input layer sequences corresponds to the size of the selected prosodic features (≤ 6). A bidirectional LSTM layer of 32 hidden units is used, followed by a ReLU (Rectified Linear Unit) activation function to allow faster and effective training of our network. We use a Dropout of 50% which role is to avoid over-fitting. Two fully connected dense layers are used followed by a Softmax.

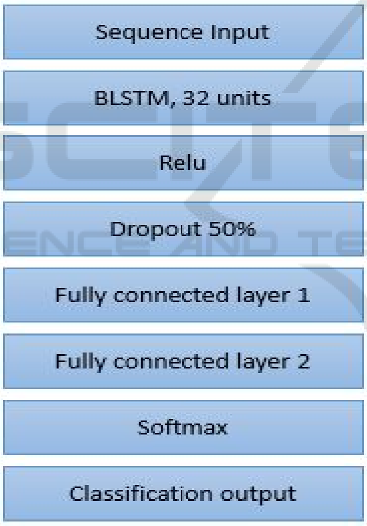

Figure 3: The proposed Long Short-Term Memory (LSTM) network architecture.

4 EXPERIMENTS AND RESULTS

4.1 Description of the RAVIOLI Database

In the RAVIOLI project, a database of injunctive and non-injunctive utterances was collected from several environments. This database is well varied over different parameters of the spontaneous speech signal (age, gender, speaker, noise, emotions, ...). The database is composed of injunctive utterances produced in authentic oral interactions collected from the ESLO2 database (http://eslo.huma-num.fr/). Each injunctive utterance is constituted of spontaneous French speech including the French word “aller” (or “allez”), which is used as key word for classifying the utterances by linguistic experts in RAVIOLI project. The use of the key word is not necessarily associated to an imperative form. The non-injunctive utterances are composed of many different speech objects like sentences, single words, interjections, murmurs, noise. The original RAVIOLI database contains about 106 hours of recording distributed in 150 sound files with length varies from 0.2 second to 5 seconds. Each audio file has a respective transcription file (obtained using the transcriber software). The RAVIOLI dataset is constituted of the four following modules:

- School: a pupils classroom (71h00)
- 24h: one day follow-up of a person (9h44)
- Itinerary: city itinerary asking and questioning (5h42)
- Meal: family or friends mealtimes (18h50)

  A fifth module of interviews were removed because of a lack of injunctive values.

4.2 Database Distribution

4.2.1 Dataset Preparation for SVM and K-NN

The RAVIOLI speech database has originally been recorded at the 44100 Hz sampling frequency. Records have been down-sampled to 16000 Hz with respect to the Shannon theory to reduce the computation cost. For the training of the two classical learning methods, the database is divided into a training dataset composed of 100 injunctive and 99 non-injunctive utterances and a testing dataset composed of 97 injunctive and 99 non-injunctive utterances.

4.2.2 Dataset Preparation for Deep Learning

Deep learning methods require a larger training dataset and the existence of a validation dataset. The validation set is used to evaluate a given model during its training phase. Hence the network occasionally sees this data, but never it learns from this. This dataset is important to fine-tune the model’s hyperparameters. The test dataset is always present as it allows to evaluate the model. It is only used once the model is completely trained. For this reason, we have changed our data configuration as follows:

- Training dataset: contains 70% of the total database corresponding to 278 audio.

Table 1: CR results given by SVM.

Configs	E	E,E_D	E,E_A	E,E_{DA}	PI	PI,PI_{DA}	PI,E	E,E_D,PI_D	PI,E,E_{DA},PI_{DA}
C,γ	10,1	0.1,10	10,1	1,1	10,1	10,0.01	10,1	1,1	10,1
CR (%)	**82**	76.02	73.46	77	63	65	65	70.91	67

Table 2: CR results given by K-NN.

Config	E	E,E_D	E,E_A	E,E_{DA}	PI	PI,PI_{DA}	PI,E	E,E_D,PI_D	PI,E,E_{DA},PI_{DA}
k	51	5	3	5	27	39	3	89	11
CR (%)	**71.42**	65.75	69.89	65.30	65.81	66.83	69.83	70.42	65.81

Table 3: CR results given by LSTM through different testing configuration of the hyper-parameters, using the 6 prosodic features.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
learning rate	0.01	0.01	0.001	0.001	0.001	0.001
momentum	0.9	0.9	0.9	0.9	0.8	0.9
mini-batch size	10	5	10	10	10	27
optimizer	ADAM	ADAM	ADAM	SGDM	SGDM	SGDM
CR (%)	74	65.5	75.64	**96.15**	80.15	88.46

Table 4: CR results given by **hyper-parameters optimized** LSTM, using increasing input sizes (number of prosodic features).

features	E	PI,E	PI,E,E_{DA},PI_{DA}
input size	1	2	6
optimizer	ADAM	SGDM	SGDM
CR (%)	79.49	80.77	**96.15**

- Validation dataset: contains 10% of the total database corresponding to 39 audio.
- Test dataset: contains 20% of the total database corresponding to 78 audio.

During the training process, by default, the program divides the training data into mini-batches and then compresses the sequences in the same batch to the same length by a zero padding. This non-optimal padding can have a negative impact on the network performance. To overcome this problem, we sorted audio in the training dataset by its duration. In this way, we allowed the model to select sequences from the same mini-batch with similar lengths thus optimising the learning process.

4.3 Classification Results and Discussion

Remember that the different configurations of descriptors are noted $TYPE_{DA}$ in which TYPE is the prosodic feature, like 'PI' for the pitch or 'E' for the log energy, 'D' is the derivative Δ (speed) and 'A' is the double derivative $\Delta\Delta$ (acceleration). The quality of the classification system is evaluated by the classification rate (CR) defined as: $CR = \frac{O-M}{O}$, where O is the total number of occurrences given at the input of the classifier and M is the number of misclassified occurrences.

4.3.1 Classification using SVM and K-NN

The first objective of this study is to confirm the previous classification results found with the GMM classifier (Hacine-Gharbi and Ravier, 2020). For this purpose, we applied the K-NN and SVM methods on the same database as described below. For K-NN, we varied the number of nearest neighbours (k all odd numbers between 1 and 200). For the SVM method, we chose the RBF kernel and we changed the values of the regularization parameter ($C = [0.01, 0.1, 0.5, 1, 1.5, 5, 10]$) and the coefficient values of the kernel ($\gamma = [0.01, 0.1, 1, 10, 100]$). Table 1 and Table 2 show the best classification rates obtained on different configurations and the optimal found parameters used for each method. From the two tables, we can note that both SVM and K-NN gave the best classification rate of 82% and 71.42% using only the logarithmic energy. This result confirms the predominance of the logarithmic energy feature regardless of the classifier used.

4.3.2 Classification using LSTM

To train a deep learning network, we need to set the hyper-parameters of the network empirically. This step is very important because it allows to decide the learning performance and the speed of a network. Thanks to the validation dataset, tuning hyper-parameters can be optimized, the total number of epochs is set to 300 for all tests, the best network throughout the training is adopted.

First, we show in Table 3 the volatility of the CR results given by LSTM through different testing configuration of the hyper-parameters. These results have been obtained considering the 6 prosodic features. By empirically optimizing the hyper-parameters set using few combinations, it is thus possible to increase the CR to the very high CR of **96.15%**.

Second, using this optimization strategy, we show CR results given by LSTM in Table 4, using increasing input sizes (number of prosodic features). A first test is to train the network using the logarithmic energy only, thus the size of the input layer sequences is equal to 1. We used the adaptive moment estimation (ADAM) as optimizer with a learning rate of 0.001 and a momentum constant equal to 0.9, to minimize the binary-cross-entropy loss function. The CR given by the trained network is 79.49%. As the input data are still small, we trained the LSTM network by using two prosodic features E, PI and the six prosodic features to increase our data collection volume. To train this new network, The ADAM optimizer doesn't give good training, we therefore used the Stochastic Gradient Descent with momentum (SGDM) optimizer, with a learning rate of 0.001 and a momentum equal to 0.9. The size of the mini-batch is 10. Using E and PI, the CR is 80.77% while using the 6 features, the CR reaches 96.15%.

5 CONCLUSIONS

In the present paper, we used two conventional classifiers, namely the support vector machine (SVM) and the k-nearest neighbors (K-NN) with a voting rule decision strategy for classifying INJ vs NINJ utterances in RAVIOLI database. By optimally tuning their parameters, these classifiers all gave the best CR values when applied with the log energy feature only compared with six features namely, the energy and the pitch with their first and second derivatives. The best classification rate of 82% was given with SVM method. When applying the Long Short-Term Memory (LSTM) network on our data, the CR reach the not better value of 79.49% by using the log energy feature alone. More surprisingly, the CR significantly increased to 96.15% by using the 6 prosodic features. Albeit more test are necessary we conclude that deep learning methods need as much data as possible for reaching high performance, even the less informative ones, especially when the dataset is small. The counterpart of deep learning methods remains the difficulty of optimal parameters tuning. Future work will investigate larger INJ and NINJ dataset extracted from RAVIOLI database as well as a injunction categorization within the injunction of the dataset.

ACKNOWLEDGEMENTS

This work is funded by the région Centre-Val de Loire, France, under the contract APR IA Ravioli #17055LLL (lotfi.abouda@univ-orleans.fr, leader). This collaborative work implies three laboratories of University of Orléans (LLL, PRISME/IRAuS, LIFO) and one laboratory from University of Tours (LIFAT). All the persons involved in this project are acknowledged for their participation.

REFERENCES

Bengacemi, H., Hacine-Gharbi, A., Ravier, P., and abed meraim, K. (2021). Surface emg signal classification for parkinson's disease using wcc descriptor and ann classifier. pages 287–294.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. version 5.3. 51. *Online: http://www. praat. org*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Fix, E. and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247.

Ghazali, F., Hacine-Gharbi, A., and Ravier, P. (2020). Statistical features extraction based on the discrete wavelet transform for electrical appliances identification. In *Proceedings of the 1st International Conference on Intelligent Systems and Pattern Recognition*, pages 22–26, Virtual Event Tunisia. ACM.

Ghazali, F., Hacine-Gharbi, A., and Ravier, P. (2021). Selection of statistical wavelet features using wrapper approach for electrical appliances identification based on knn classifier combined with voting rules method. *International Journal of Computational Systems Engineering*, page In Press.

Ghazali, F., Hacine-Gharbi, A., Ravier, P., and Mohamadi, T. (2019). Extraction and selection of statistical harmonics features for electrical appliances identification using k-nn classifier combined with voting rules method. *Turkish Journal of Electrical Engineering & Computer Sciences*, 27:2980–2997.

Hacine-Gharbi, A., Petit, M., Ravier, P., and Némo, F. (2015). Prosody based Automatic Classification of the Uses of French 'Oui' as Convinced or Unconvinced Uses:. In *Proceedings of the International Conference on Pattern Recognition Applications and Meth-*

ods, pages 349–354, Lisbon, Portugal. SCITEPRESS - Science and and Technology Publications.

Hacine-Gharbi, A. and Ravier, P. (2020). Automatic classification of french spontaneous oral speech into injunction and no-injunction classes. In *ICPRAM*, pages 638–644.

Hacine-Gharbi, A., Ravier, P., and Nemo, F. (2017). Local and Global Feature Selection for Prosodic Classification of the Word's Uses:. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, pages 711–717, Porto, Portugal. SCITEPRESS - Science and Technology Publications.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1999). The htk book (version 2.2). entropic ltd., cambridge.