Offline Speech Recognition Development
A Systematic Review of the Literature

Lucas Debatin[1,2], Aluizio Haendchen Filho[2] and Rudimar L. S. Dazzi[1]
[1]*Laboratory of Applied Intelligence, University of Vale do Itajaí (UNIVALI), Itajaí, SC, Brazil*
[2]*Department of Artificial Intelligence and Smart Systems, University Center of Brusque (UNIFEBE), Brusque, SC, Brazil*

Keywords: Voice Recognition, Offline Recognition, Mobile Devices.

Abstract: This paper aims to present the state-of-the-art of speech recognition from a systematic review of the literature. For this, 222 papers from four digital repositories were examined. The research followed a methodology composed of questions of search, expression of search and criteria of inclusion and exclusion. After reading the abstract, introduction and conclusion, nine papers were selected. Based on the analysis of the selected papers, we observed that the research prioritizes the following topics: (i) solutions to reduce the error rate; (ii) neural networks for language models; and (iii) n-gram statistical models. However, no solution was offered to provide offline voice recognition on Android mobile devices. The information obtained is very useful in order to acquire knowledge to be used in the development of offline voice recognition in mobile devices. The techniques provide guidelines for the application of the best neural networks and mechanisms for reducing error rates.

1 INTRODUCTION

The speech recognition procedure aims to convert spoken language into text. This facilitates and makes communication more natural for humans, since the purpose of this recognition is to enable people to communicate more naturally and effectively (Huang and Deng, 2009).

Continuous speech recognition is complex and challenging to implement, since it must be able to handle all the characteristics and voices of language in the natural way, such as unknown word lengths, coarticulation effects and sloppy pronunciation (Alencar, 2005). This can be solved using language models (n-gram) to find optimal estimates for conditions-conditioned probabilities (Silva et al., 2004).

Some factors contribute to the demand for adaptive interfaces with speech recognition. First, because it is a natural way to use speech, and second, because touch interfaces makes it difficult to write long texts (Hearst, 2011). Thus, the use and access to information in software and applications can becomes faster, since the keyboard, mouse and touch-screen are not used as input methods.

Advances in speech recognition techniques enable the use of this technology in many applications, particularly mobile devices. People with special needs are also benefited by such systems. Users who cannot use their hands and visually impaired people use this technology to express themselves by performing control over various computer functions through voice (Silva, 2010).

There are several Application Programming Interface (API) that facilitate the implementation of voice recognition in software and applications, such as Web speech, Java speech, Google cloud speech, Bing speech, among others. However, none of them perform recognition in offline mode, that is, the user must be connected to the Internet. This limitation is a great barrier, because in countries like Brazil, for example, only approximately 50%of the population has access to the internet (IBGE, 2014). This is a great concern, considering that voice recognition is an important means of accessibility.

Another limitation is that many of the current APIs for online voice recognition are proprietary software, that is, they are not free. In many cases, the amount paid becomes high, as it depends directly on the number of requests that the API performs. Voice recognition is also widely used in enterprise software and, in many cases, it is necessary to be offline and free.

Debatin, L., Haendchen Filho, A. and L. S. Dazzi, R.
Offline Speech Recognition Development.
DOI: 10.5220/0006788005510558
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 551-558
ISBN: 978-989-758-298-1

Offline speech recognition presents high computational complexity, and requires a large amount of memory (Alencar, 2005). However, this requirement is not a very limiting feature, since today smartphones are increasingly modern and powerful. For offline recognition, it will be necessary to use a neural network and statistical model, which together have a good processing and memory usage index, so it preserves the smartphone's battery consumption.

This paper presents a systematic review of the literature, and analyzes the state of the art on the development of offline continuous speech recognition for Android mobile devices. The review was carried out selecting papers published in the last five years in four different repositories: ACM, IEEE, ScienceDirect and Scopus.

In this context, this work aims to identify the resources used for speech recognition that can significantly contribute to the reduction of error rate. This stage mainly involves: (i) identifying the most relevant solutions according to the inclusion and exclusion criteria defined in the methodology; (ii) to know the different techniques used; and (iii) to analyze the main advantages and disadvantages of these techniques.

This paper is divided into the following sessions: (i) methodology, which presents the criteria used to carry out the systematic review; (ii) background; (iii) results; (iv) discussion, establishing relationships between the various results and their general implications for the problem addressed; and (v) conclusions and future work.

2 METHODOLOGY

According to Kitchenham and Charters (2007), a systematic literature review is a means of identifying, evaluating, and interpreting all available and relevant research for a particular research question, or subject, or phenomenon of interest.

The most common reasons for undertaking a systematic review are: to summarize existing evidence on the treatment of technology; identify gaps in current research; and provide a background for appropriately positioning new research activities (Kitchenham and Charters, 2007).

The methodology applied in this systematic review of the literature consists of delimiting: (i) the research questions; (ii) repositories and research strategy; and (iii) selection of articles.

2.1 Research Questions

Based on the objective of this systematic review, four research questions were formulated, which are presented in Table 1.

Table 1: Research questions.

ID	Research question
P1	What neural networks, within this area of voice recognition, are being researched more?
P2	What solutions are being studied to reduce voice recognition error rates?
P3	Is the n-gram statistical model used to improve speech recognition?
P4	What ways to make offline voice recognition available on Android mobile devices?

2.2 Repositories and Research Strategy

To answer these questions, four different electronic repositories of research were selected. In Table 2, the name and web access address of each can be seen.

Table 2: Electronic repositories.

Repository	Access Address
ACM	http://www.acm.org
IEEE	http://ieeexplore.ieee.org
ScienceDirect	http://www.sciencedirect.com
Scopus	https://www.scopus.com

Based on the selected repositories, a search expression was developed from keywords that did not present redundancy in the results. The goal is to contemplate the largest number of articles, and at the same time act as a filter to return the most relevant papers on the subject.

We used the following search expression, used in the search in each repository: (("speech recognition" **OR** "offline speech recognition" **OR** "continuous speech recognition" **OR** ((mobile **OR** android) **AND** "speech recognition")) **AND** ("neural network" **OR** "deep learning") **AND** ("n-gram" **OR** ("natural language processing" **OR** nlp))). When analyzing this expression, it is possible to notice that keywords were used that refer to the problem of the study.

2.3 Papers Selection

The inclusion and exclusion criteria for choosing articles are presented in Table 3. Exclusion criteria CE3 and CE4 were used to select only the most recent and relevant papers on the theme.

Table 3: Inclusion and exclusion criteria.

Inclusion	Exclusion
CI1: papers published between 01/01/2012 and 06/30/2017. CI2: search expression by filtering the papers by title, abstract, and keywords. CI3: papers in English and Portuguese.	CE1: short papers (expanded abstracts) CE2: conferences without a public review committee. CE3: papers that have less than five citations. CE4: papers with more than one year of publication that have less than ten citations. CE5: articles that do not present the use of neural networks.

In addition to the criteria, the papers were also selected by reading the following topics: (i) title and keywords used; (ii) summary; and (iii) introduction and conclusion. These criteria for reading and selection were helpful in minimizing the effort in reading and selecting papers that actually contribute to answering the research topic questions.

3 BACKGROUND

In this chapter will be presented concepts and characteristics of the neural networks and n-gram statistical model, used in the development of voice recognition. This background contributes to a better understanding of the results described in Section 4.

3.1 Artificial Neural Networks

The ability to computationally implement simplified versions of biological neurons has given rise to a subspecialty of Artificial Intelligence, known as Artificial Neural Networks (ANNs), which can be defined as parallel systems composed of simple, layered and highly interconnected, inspired in the human brain (Haykin, 2001).

Engel (2002) defines ANNs as distributed parallel processors of large capacity, consisting of simple processing units, which are interconnected by links that have adjustable parameters. These adjustable parameters were called synaptic weights, in analogy to the biological synapses, and allow to control the intensity of the connections between the neurons that form the networks.

According to Tsai et al. (2001), the network operation is largely determined by the value of these connections (weights) among its elements. The synaptic weights store the knowledge represented in the model and serve to weight the input received by each neuron in the network.

ANNs differ by their architecture, by how the weights associated with the connections are adjusted during the learning process, by the type of training and the activation function used. The architecture of a neural network restricts the type of problem in which the network can be used, and is defined by the number of layers (single layer or multiple layers), number of nodes in each layer and type of connection between nodes (Haykin, 2001).

Next, the ANNs models obtained through the selected articles are related, and the concepts and characteristics of the implementation of these models in the voice recognition are presented.

3.1.1 DNN

A DNN (Deep Neural Network) consists of a succession of convolutional, max-pooling and fully connected layers. It is a general, hierarchical feature extractor that maps raw pixel intensities of the input image into a feature vector to be classified by several fully connected layers. All adjustable parameters are jointly optimized through minimization of the misclassification error over the training set (Ciresan et al., 2012).

They are successfully applied in acoustic models of state-of-the-art speech recognition systems. It is a set of hidden layers with linear transformations and nonlinear activations to make predictions. This approach allows complex data to be well modeled (Dahl et al., 2012).

3.1.2 NN-LM

NNLM (Neural Network for Language Modeling) solves the problem of dimensionality suffered by n-gram models and allows the continuous representation of words (Bengio et al., 2003). This model uses neural networks in conjunction with statistical models.

The idea of discovering some similarities between words to obtain generalization from training sequences to new sequences is not new. For example, it is exploited in approaches that are based on learning a clustering of the words (Brown et al., 1992, Pereira et al., 1993, Niesler et al., 1998, Baker and McCallum, 1998): each word is associated deterministically or probabilistically with a discrete class, and words in the same class are similar in some respect.

3.1.3 RNN-LM

The simple RNN-LM (Recurrent Neural Network for Language Modeling) consists of an input layer, a hidden layer with recurrent connections that propagate time-delayed signals, and an output layer, plus the corresponding weight matrices. The input vector w(t) represents input word at time t encoded using 1-of-N coding (also called one-hot coding), and the output layer produces a probability distribution over words. The hidden layer maintains a representation of the sentence history. The feature layer represents an external input vector that should contain complementary information to the input word vector w(t) (Mikolov et al., 2010).

This neural network improves and reduces error rates in speech recognition systems compared to traditional n-gram approaches (Mikolov et al., 2010).

3.1.4 BRNN-LM

BRNN-LM (Bayesian Recurrent Neural Network for Language Modeling) is a Bayesian learning method for RNN-LM where a regularization term was introduced to conduct the penalized model training (Chien and Ku, 2016).

It compensates for the uncertainties of the model by minimizing the regularized entropy error function for a classification network, in which the regularization term is governed by a Gaussian parameter (Chien and Ku, 2016).

3.1.5 HMM-ANN

Hybrid HMM/ANN (Hidden Markov Model - Artificial Neural Networks) models compute the emission probabilities for the HMMs with a neural network instead of the commonly used Gaussian mixtures. These emission probabilities are provided with a neural network since ANNs can be trained to estimate probabilities. The estimates of the posterior probabilities computed by the neural network are divided by the prior state probabilities, resulting in scaled likelihoods (Espanã-Boquera et al., 2011).

There are two advantages over the probability indexes for calculating confidence estimates: (i) no additional models are required to normalize the output of the acoustic model because they are automatically normalized with the discriminatory training criterion; and (ii) the acoustic model is trained according to the MAP (Maximum A Posteriori) criterion, which is naturally discriminatory and may be preferred over the MMI criterion (Williams and Renals, 1997).

3.1.6 LSTM

The LSTM (Long Short-Term Memory) architecture contains special units called memory blocks in the recurring hidden layer. The memory blocks contain cells with self-connections storing the temporal state of the network, as well as special multiplicative units, called gateways to control the flow of information (Beaufays et al., 2014).

In particular, the LSTM architecture, which overcomes some modeling weaknesses of RNNs, is conceptually attractive for the task of acoustic modeling (Beaufays et al., 2014).

3.1.7 MTL-DNN

MTL (Multi-Task Learning) is a machine learning approach that aims to improve the generalization performance of a learning task with the use of many related tasks. MTL has been successfully applied to many speech, language, image and vision tasks using the neural network (NN) because its hidden layers naturally capture learned knowledge that can easily be transferred or shared across multiple tasks (Chen and Mak, 2015).

3.2 Natural Language Processing

Natural language is the same language that human beings use daily to communicate with one another. To understand this language, the computational system must be able to process and manipulate it on several levels. This is the goal of natural language processing, which is a subarea of Artificial Intelligence.

According to Coppin (2010), Artificial Intelligence involves the use of methods based on the intelligent behavior of humans to solve complex problems.

The five levels of processing and manipulation of language, according to Coppin (2010), are: (i) phonology; (ii) morphology; (iii) syntax; (iv) semantics; and (v) pragmatic. In addition, natural language processing is essentially focused on three aspects of natural language communication: sound (phonology), structure (morphology and syntax) and meaning (semantic and pragmatic).

In addition to the above-mentioned levels, natural language processing should apply some world knowledge, i.e. to know real-world subjects, since the purpose of natural language processing would be to have a system with enough world knowledge to be able to engage in a discussion with humans on any subject (Coppin, 2010).

Next, the concepts and characteristics of the n-gram back-off statistical model, which was obtained through the selected articles, will be commented.

3.2.1 Back-off

In an n-gram language model, we treat two histories as equivalent if they end in the same n - 1 words, i.e., we assume that for $k \geq n$, Pr ($\omega_k \mid \omega_1^{k-1}$) is equal to $\omega_k \mid$ Pr (ω^{k-1}_{k-n+1}). For a vocabulary of size V, a 1-gram model has V - 1 independent parameters, one for each word minus one for the constraint that all of the probabilities add up to 1. A 2-gram model has $V(V$ - 1) independent parameters of the form Pr ($\omega_2 \mid \omega_1$) and V - 1 of the form Pr (ω) for a total of V^2 - 1 independent parameters. In general, an n-gram model has V^n - 1 independent parameters: V^{n-1} (V - 1) of the form Pr ($\omega_n \mid \omega^{n-1}_1$), which we call the order-n parameters, plus the V^{n-1} - 1 parameters of an (n - 1)-gram model (Jurafsky and Martin, 2008).

The back-off model uses trigram if the evidence is sufficient, otherwise it uses bigram or unigram. In other words, it only retreats to a lower order n-gram if it has zero evidence for a top-level interpolation n-gram (Jurafsky and Martin, 2008).

4 REVISION RESULTS

When applying the search expression and using the inclusion and exclusion criteria in each repository, 222 papers were discovered. Then, the selection of these papers was made, analyzing the title, keywords, summary, introduction and conclusion. After this selection, the number of papers was reduced to 9, that is, only 4% of the articles discovered showed some strong direct relation with the proposed theme. Table 4 shows the number of papers discovered and selected per repository.

Table 4: Number of articles discovered and selected.

Repository	Discovered	Selected
ACM	21	0
IEEE	17	4
ScienceDirect	147	2
Scopus	37	3

Table 5 presents the selected papers, each with its reference (used to identify the papers), number of citations (up to June 30, 2017) and year. Number of citations sorts the papers.

Figure 1 shows a graph with the papers selected by year. Most of these papers were published in the year 2013. Many recent papers were not selected because they were not relevant due to the CE3 and CE4 exclusion criteria.

Table 5: List of selected articles.

Reference	Citations	Repository	Year
(Sundermeyer et al., 2015)	49	IEEE	2015
(Arisoy et al., 2014)	33	IEEE	2014
(Deoras et al., 2013)	25	Scopus	2013
(Siniscalchi et al., 2013)	19	Scopus	2013
(Liu et al., 2013)	18	ScienceDirect	2013
(Shi et al., 2013)	16	Scoupus	2013
(Chen and Mak, 2015)	15	IEEE	2015
(Chien and Ku, 2016)	10	IEEE	2016
(Maas et al., 2017)	6	ScienceDirect	2017

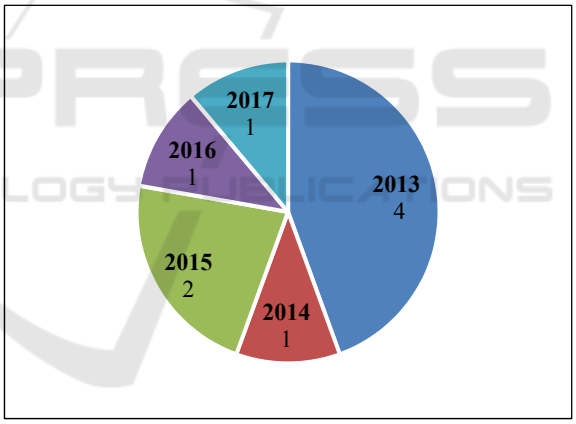

Figure 1: Selected articles by year.

4.1 Papers Analysis

Table 6 presents the selected papers and the first three research questions. The questions were answered by reading each article, in a synthesized way, in order to facilitate the interpretation and tabulation of the data. The fourth question, related to the use of offline voice recognition, does not appear in the frame, as no article has confirmed the use of this procedure. In this table are all the articles that form the final result of the systematic literature review.

It was verified that the main neural network used is NN-LM (Neural Network Language Modeling).

For each neural network found, a brief study was carried out to identify the main characteristics of these networks in the implementation of speech recognition (see topic 3.1).

Table 6: Research question answers by paper.

Reference	Neural Network	How it reduces the error rate	N-gram
(Sundermeyer et al., 2015)	LSTM	A single two layers LSTM	Yes
(Arisoy et al., 2014)	NN-LM	Method for approaching an NN-LM with a back-off LM	Yes, back-off
(Deoras et al., 2013)	NN-LM	Using a variational model	Yes
(Siniscalchi et al., 2013)	HMM-ANN	Combining the attribute scores with the HMM probability scores	Yes
(Liu et al., 2013)	NN-LM	ROVER process and acoustical cross-matching model	Yes
(Shi et al., 2013)	RNN-LM	Using the RNN-LM-Brown	Yes, back-off
(Chen and Mak, 2015)	MTL-DNN	Two methods in the multitasking learning structure	No
(Chien and Ku, 2016)	BRNN-LM	Bayesian recurrent neural network for language modeling	Yes
(Maas et al., 2017)	DNN	DNN architecture and an optimization technique	No

In addition, it can be seen that 77% of the papers use the n-gram statistical model, and the main model used is the back-off. The others do not use n-gram to improve speech recognition error rates.

It was also verified that all the selected papers presented some solution to reduce the rates of voice recognition errors. In many cases, it can be seen that the solution is associated with the joint use of neural networks and n-gram statistical models. In addition, no paper has provided any way to make voice recognition available offline.

5 DISCUSSION

The selected papers were very useful in order to acquire knowledge to be used in the development of off-line voice recognition in mobile devices. The techniques provide guidelines for the application of the best neural networks and mechanisms for reducing error rates.

The use of the inclusion and exclusion criteria defined in the methodology helped to select the most relevant articles and were useful in finding the best solutions. All articles presented answers and solutions to the research questions, with different alternatives for solving the problem.

We observed that there is great diversity of artificial neural types that there is great diversity of artificial neural networks types used in the development of voice recognition solutions. Each one has peculiar characteristics related to the resolution of the problem. An important aspect detected was the confirmation that the NN-LM model is the most frequently used method to solve this problem.

We also concluded that the reduction of the error rate is associated with the joint use of artificial neural networks and n-gram techniques. The n-gram back-off statistical model stands out in the approaches that use NN-LM, as verified by Arisoy et al. (2014) and Shi et al. (2013).

One the aspects analyzed is the differences between the variants of the NN-LM model that are RNN-LM of Shi et al. (2013) and BRNN-LM of Chien and Ku (2016).

Neural networks offer considerable improvements when interpolated with LM, but are overcome by RNN-LM, as they show additional reductions in perplexity and word error rate (WER) (Sundermeyer et al., 2015). It is important to note that perplexity improvement is a precise predictor to reduce error rates (Sundermeyer et al., 2015).

On the other hand, BRNN-LM presented a better consistency in relation to NN-LM in perplexity and WER (Chien and Ku, 2016). Using a NN-LM to construct a backoff-gram model, it is possible to achieve a WER improvement of up to 1% absolute, without compromising decoding speed and without the need to modify the existing decoding software (Arisoy et al., 2014).

In addition, when comparing LSTM with RNN-LM in English development data, there is an additional perplexity reduction of 14%. The LSTMs

still present more consistent improvements when adding hidden layers, since with a single two-layer LSTM, the error rate was improved from 12.4% to 10.4% (Sundermeyer et al., 2015). Increasing the size and depth of the DNN model are simple but effective ways of improving WER performance, but only to a limited extent, because raising the depth too much can worsen its performance (Maas et al., 2017).

The MTL-DNN model is suitable for training phonetic models of low-resource languages, such as some dialects. This is possible because both the hidden inputs and layers are shared by the multiple learning tasks. A disadvantage is that multitasking learning, while being a powerful learning method, requires that all the tasks involved are truly related (Chen and Mak, 2015).

Regarding the HMM-ANN model, it was found that a conventional decoupled HMM-based system can not provide the required precision, which resulted in dramatically degraded accuracy. For this reason, it is necessary to combine the acoustic scores of the WFSM system with the acoustic scores of a system based on MMI HMM trained discriminatorily on the best lists included in the network of words generated by the system (Siniscalchi et al., 2013).

Offline speech recognition processing may have problems with smartphones, such as high battery consumption, slower running in the background, and other restrictions. Knowing the advantages and gains of each approach becomes important for selecting an ANN that performs well in memory usage and has a lower error rate for offline speech recognition, and the lowest possible battery consumption.

We did not find any approach for offline speech recognition in literature. In this sense, the work will be challenging and at the same time innovative, presenting an unprecedented research theme.

6 CONCLUSIONS AND FUTURE WORK

Online voice recognition essentially depends on the Internet. Sometimes this feature may be unavailable in many places, particularly in those that do not have unlimited Internet. There is also a tendency for this technology to become unavailable. This may lead to restrictions in businesses and for the public, because there will be related costs depending on the time utilization. In addition, for online voice recognition, a large processing capacity is required, in which API developers have to invest in their servers, as several users request audio processing at the same time. Therefore, it may be important to have this feature offline in mobile devices.

With the information and knowledge gained from this research, future work will focused on the design and implementation of offline voice recognition for mobile devices. The next step will be to test the performance of n-gram and neural network models in this context. A comparative performance analysis will be performed through the implementation of the main neural networks and statistical models used in the continuous voice recognition, which were found in the systematic review of the literature.

The remaining steps will be based on the choice of neural networks and statistical models with the best performance, and will serve to: (i) develop an Android mobile application to investigate the processing and memory usage in various smartphones; and (ii) to verify the error rate for Brazilian Portuguese.

At the end of these steps, we intend to achieve the goal of developing a solution with a lower processing rate, minimizing errors and memory use, hence reducing smartphone's battery consumption.

ACKNOWLEDGEMENTS

We thank CAPES (Coordination of Higher Education Student Improvement) for financial support to carry out this research work.

REFERENCES

Alencar, V. F. S. (2005) “Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído”, In: Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

Arisoy, E., Chen, S. F., Ramabhadran, B. and Sethy, A. (2014). Converting Neural Network Language Models into Back-off Language Models for Efficient Decoding in Automatic Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Baker, D. and McCallum, A. (1998). Distributional clustering of words for text classification. SIGIR’98.

Beaufays, F., Sak, H., and Senior, A. (2014). Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling Has. Interspeech.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). A Neural Probabilistic Language Model. Proc. J. Machine Learning Research.

Brown, P.F., Della Pietra, V.J., DeSouza, P.V., Lai, J.C. and Mercer, R.L. (1992). Class-based n-gram models of natural language. Computational Linguistics.

Chen, D. and Mak, B. K. (2015). Multitask Learning of Deep Neural Networks for Low-Resource Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Chien, J. and Ku, Y. (2016). Bayesian Recurrent Neural Network for Language Modeling. IEEE Transactions on Neural Networks and Learning Systems.

Ciresan, D. C., Meier, U. and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. Computer Vision and Pattern Recognition.

Coppin, B., 2010. *Inteligência Artificial*, LTC. Rio de Janeiro, 1nd edition.

Dahl, G. E., Yu, D., Deng, L. and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing.

Deoras, A., Mikolov, T., Kombrink, S. and Church, K. (2013). Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model. Speech Communication.

Engel, P. M. (2002) "Redes Neuais – Notas de aula", In: Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

Espanã-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J. and Zamora-Martinez, F. (2011). Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Haykin, S. 2001. *Redes Neurais: Princípios e prática*, Bookman. Porto Alegre, 2nd edition.

Hearst, M. A. (2011). 'Natural' search user interfaces. Communications of the ACM.

Huang, X. and Deng, L. (2009). An Overview of Modern Speech Recognition. Microsoft Corporation.

IBGE. (2014) "Pesquisa Nacional por Amostra de Domicílios", https://goo.gl/7uVigP, Junho 2017.

Jurafsky, D. and Martin, J. H. (2008), Speech and Language Processing, Prentice Hall, 2th edition.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report.

Liu, X., Gales, M. J. F. and Woodland, P.C. (2013). Language model cross adaptation for LVCSR system combination. Computer Speech & Language.

Maas, A. L., Qi, P., Xie, Z., Hannun, A. Y., Lengerich, C. T., Jurafsky, D. and Ng, A. Y. (2017). Building DNN acoustic models for large vocabulary speech recognition. Computer Speech & Language.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. and Khu-danpur, S. (2010). Recurrent neural network based language model. Proceedings of Interspeech.

Niesler, T. R., Whittaker, E. W. D. and Woodland, P.C. (1998). Comparison of part-of-speech and automatically derived category-based language models for speech recognition. International Conference on Acoustics, Speech and Signal Processing.

Pereira, F., Tishby, N. and Lee, L. (1993). Distributional clustering of english words. Annual Meeting of the Association for Computational Linguistics.

Shi, Y., Zhang, W., Liu, J. and Johnson, M. T. (2013). RNN language model with word clustering and class-based output layer. Journal on Audio, Speech, and Music Processing.

Silva, C. P. A. (2010). "Um software de reconhecimento de voz para português brasileiro". In: Universidade Federal do Pará, Pará, Brasil.

Silva, E., Pantoja, M., Celidônio, J. and Klautau, A. (2004) "Modelos de Linguagem N-grama para Reconhecimento de Voz com Grande Vocabulário", In: Laboratório de Banco de Dados, Belo Horizonte, Brasil.

Siniscalchi, S. M., Svendsen, T. and Lee, C. (2013). A Bottom-Up Modular Search Approach to Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing.

Sundermeyer, M., Ney, H. and Schlüter, R. (2015). From Feedforward to Recurrent LSTM Neural Networks for Language Modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Tsai, C. D., Wu, T. T., Liu, Y. H. (2001). Application of Neural Networks to Laser Ultrasonic NDE of Bonded Structures. NDT&E Internacional.

Williams, G. and Renals, S. (1997). Confidence measures for hybrid HMM/ANN speech recognition. Proc. Eurospeech.