BUILDING A LVCSR SYSTEM FOR ROMANIAN: METHODS AND CHALLENGES

Paula Georgiana ZĂLHAN
Babeş-Bolyai University of Cluj-Napoca, Faculty of Economic Sciences and Business Administration
Cluj-Napoca, Romania
paula.zalhan@econ.ubbcluj.ro

Abstract*: The aim of this paper is to present a brief survey in the field of Automatic Speech Recognition (ASR) and major advances made in the last decades of research in order to highlight the fundamental progress that has been made so far. After years of development, the accuracy of Large Vocabulary Continuous Speech Recognition (LVCSR) systems remains the most important research challenge due to several degrading factors such as variations of the context, speakers, and environment. Various methods adopted by research workers when building Large Vocabulary Speech Recognition systems are presented based on the architectural components of such systems. Challenges existing in LVCSR for Romanian language and various techniques to deal these challenges have been presented in chronological order.*
Keywords*: Large Vocabulary Continuous Speech Recognition, Feature Extraction, Acoustic Modelling, Language Model, Hidden Markov Model, Speaker Adaptation*

1. INTRODUCTION

In past years, there has been a significant growth in the field of ASR for high-resource languages. Researches in this area have been focused on the most spoken languages of the world (such as English, for example) for which there is a notable commercial success of ASR systems. ASR systems such as: systems for name-dialing (Suontausta et al. 2000; Gao et al. 2001), travel reservations (Pellom et al., 2001), getting weather- information (Zue et al., 2000), accessing financial accounts (Davies et al., 1999), automated directory assistance (Jan et al., 2003), and dictation systems (Wegmann et al., 1996; Saon et al., 2003) are already integrated in powerful solutions that are on the market.

The main advantage of these languages is the fact that speech resources needed to build acoustic models and linguistic resources needed to build general or domain-specific language models are widely available. On the opposite side, due to lack of such resources, some languages have received much less attention.

For Romanian, as under-resourced language (Berment, 2004; Cucu et al., 2014), there are just a few freely usable speech and linguistic resources such as transcribed and annotated speech corpora, phonetic dictionaries. Even though several research groups have been focused on creating such resources, these are not publicly made available. In this way, a LVCSR system for Romanian has not been developed yet. This is the main reason which calls for further research in this domain.

There are some important aspects that need to be taken into consideration when building a LVCSR system for Romanian language. Firstly, specifics of Romanian become a challenge for computational processing of the language. Romanian is a highly inflected language with various linguistic particularities (Trăndăbăț et al., 2012) that influence the acoustic process modeling. Secondly, developing a high accuracy LVCSR system is a difficult task. Variability in speech due to several degrading factors such as speaker's vocal tract, environment characteristics or transducer type become a research challenge in building speech recognition systems.

The remainder of the paper is organized as follows: Section2 provides the ASR classification based on variation parameters that influence the ASR task. Then, in Section 3 is described the mathematical formulation based on Bayesian model of a speech recognizer. Architectural details and main algorithms of each component of a LVCSR system are explained in Section 4. Section 5 presents national effort in building speech recognition systems for Romanian language. The last section concludes the paper.

2. ASR CLASSIFICATION

ASR systems can be classified according to some variation parameters that are related to the task as it is illustrated in Figure 1. Type of speech or speaking mode, dependence on speaker, size of vocabulary and bandwidth are the different basis on which researchers have worked (Ghai and Singh, 2012). Some of these parameters are:

-vocabulary size: one dimension of variation in speech recognition tasks is the vocabulary size. Speech is easier when the vocabulary to recognize is small or medium. For example, the task with two word vocabulary, like recognizing "yes" or "no" words, or the task with ten word vocabulary, like recognizing digits, are relatively easier when compared to large vocabulary tasks. Tasks like transcribing broadcast news or telephone conversations involve vocabularies of thousands of words and are much harder to handle (Adami, 2010).

-speech mode: A second dimension of variation is how natural the speech is and determines what types of utterances the speech recognition system is able to recognize. In isolated word recognition (such as digit recognition) or connected words recognition (such as sequence of digits recognition) each word is surrounded by pause. These types of recognition are much easier than continuous recognition, where users speak almost naturally and the ASR system has to determine the utterance boundaries.

-speaking style: The level of difficulty of continuous speech recognition task depends on the type of interaction. Recognizing speech from human-human interactions (such as spontaneous speech recognition by transcribing business meetings conversations, telephone conversations or broadcast news) is more difficult than recognizing speech from human-machine interactions (such as read speech which simulates the dictation task) (Jurafsky and Martin, 2009).

-speaker mode: Another dimension of variation is speaker characteristics (regional accent, gender, speaking rate, vocal effort etc.). There are two types of speech recognition systems regarding the speaker mode: speaker dependent and speaker independent systems. Speaker dependent systems can be used by a specific speaker; meanwhile

speaker independent systems can deal with different accents and pronunciations (Rabiner et al., 1979).
-transducer type: The type of device used to record the speech can affect the speech signal. The recording may rage from high-quality,head-mounted microphones to cell phones.
-background noise: A final dimension of variation is noise. Noise of any kind makes the recognition harder. Thus recognizing a speaker dictating in a quiet office is much easier that recognizing a speaker dictating in noisy environments.

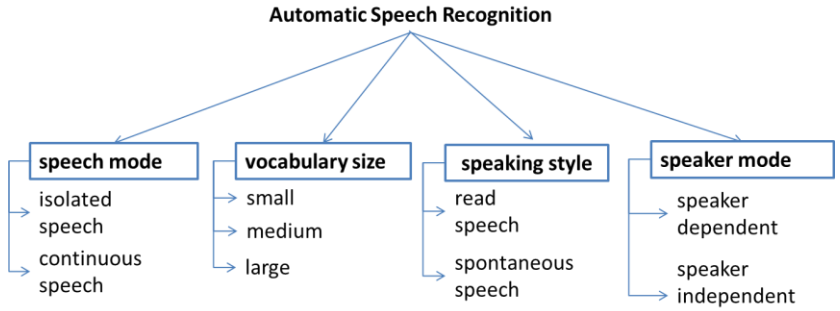

Figure 1 ASR classification based on some variation parameters related to the task

3. MATHEMATICAL FORMULATION OF LVCSR

The standard approach to LVCSR is to assume the probabilistic or Bayesian model whereby a speech signal corresponds to a word or sequence of words, in the vocabulary, with a certain probability. The input speech signal from a microphone is converted into a sequence of acoustic observations $O = o, o_2, \dots, o_n$ in a process of feature extraction. Assuming that a sequence of words $W = w_1, w_2, \dots, w_n$ was spoken, the decoder computes a probability for matching these words with given acoustic observations$O$. Finally, from all possible sequences of words, the decoder selects the one with highest probability. This sequence of words with highest probability is more likely to be produced given the observed acoustic evidence.

The implementation of the probabilistic model described above can be expressed as follows:

$$\hat{W} = \underset{W}{\boldsymbol{arg\,max}}\, \boldsymbol{P(W|O)} \quad (1)$$

Using Bayes' rule, Equation (1) can be rewritten as:

$$\hat{W} = \underset{W}{\boldsymbol{arg\,max}}\, \frac{\boldsymbol{P(W)P(O|W)}}{\boldsymbol{P(O)}} \quad (2)$$

Since $P(O)$ is independent of W, the equation (2) is equivalent to:

$$\hat{W} = \underset{W}{\boldsymbol{arg\,max}}\, \boldsymbol{P(O|W)P(W)} \quad (3)$$

The first term in equation (3), $P(O|W)$ is generally called the acoustic model (AM), as it estimates the probability of a sequence of acoustic observations *O* when a speaker utters a sequence of words *W*. Thus, $P(O|W)$ is closely related to phonetic modeling. The probability $P(O|W)$should take into account speaker variations, pronunciation variations, environmental variations, and context-dependent phonetic co-articulation variations.

The second term in equation (3), $P(W)$ is called the language model (LM) and expresses how likely a given sequence of words *W* is to be the source sentence uttered. The LM is typically an*n*-gram model in which the probability of each word is conditioned only on its *n-1*predecessors.

Given the AM and LM probabilities, the probabilistic model can be operationalized in a search algorithm that tries to generate the word sequence that has the maximum probability for a given acoustic waveform (Jurafsky and Martin, 2009).

4. ARCHITECTURE OF LVCSR

The principal components of a LVCSR system are illustrated in Figure 2. A LVCSR system takes as input the speech from a recorded audio signal and produces as output the sequence of words corresponding to the input speech signal.

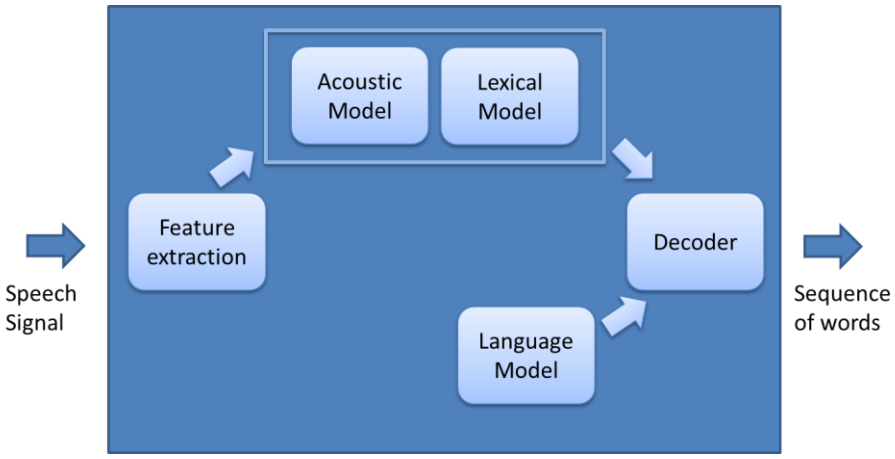

Figure 2 LVCSR system components

4.1. FEATURE EXTRACTION

Feature extraction or signal processing requires much attention because the performance of an LVCSR system depends heavily on this stage. Feature extraction seeks to provide a compact representation of the speech waveform. This representation should compress the speech data while keeping the linguistic information, despite the speaker, channel or environment characteristic. Thus, the acoustic waveform is sampled into frames (usually of 10, 15, or 20 milliseconds), each frame having a duration of 20 to 25 milliseconds. These frames are then transformed into a sequence of acoustic features. The

most used methods for feature extraction are Mel Frequency Cepstral Coefficient (MFCCs) (Davis and Mermelstein, 1980) and Perceptual Linear Prediction (PLP) (Hermansky, 1990). MFCCs are the results of a cosine transformation of a log of short-term power spectrum expressed on a non-linear Mel scale of frequency. Although this method is proven to be more efficient than other methods (Tiwari, 2010), PLP can give better results than MFCCs, especially in noisy environments (Young, 2008). PLP computes linear prediction coefficients applying several psychophysically-based spectral transformations on the short-term power spectrum and then transforms the linear prediction coefficients to cepstral coefficients.

4.2. ACOUSTIC MODEL

The acoustic model is the main component of a speech recognition system and is developed in order to establish a statistical representation for the feature vector sequences computed from the input speech. This model computes the probability of the observed feature vectors given linguistic units (words, phones). It is responsible for detecting the spoken phoneme which is defined as the smallest unit of speech that distinguishes a meaning according to (Gruhn, Minker and Nakamura, 2011).

For LVCSR it is important to decompose a word into sub-word speech units (such as phones) and build statistical models for these sub-word speech units. For every spoken word, the corresponding acoustic model is built by concatenating phoneme models to make words as defined by a pronunciation dictionary.

The most effective type of AM model is Hidden Markov Model (HMM) (Baum, 1972). According to HMM based AM, it is assumed that the feature vector sequences corresponding to each word is generated by a Markov chain. Thus, the speech is modeled as a sequence of states, where each state of the Markov chain corresponds to a single phone. In such model, a word HMM consists of a sequence of HMM states concatenated together. What it is important to mention here is that there are not allowed arbitrary transitions between states due to sequential nature of speech; states can only transition to themselves or to successive states (Jurafsky and Martin, 2009).

In order to compute for each HMM state the likelihood of a given feature vector, there are several classifiers used to estimate the AM model. For example, Subspace Gaussian Mixture Models (SGMMs) have been proposed to deal with under-resourced languages and recent studies (Povey et al., 2011) show that SGMMs can outperform the conventional Gaussian Mixture Models (GMMs), particularly with smaller amounts of training data which consist of audio recordings of speech and their text scripts. Due to SGMMs' limitations (Povey et al., 2011), most speech recognition systems use GMMs as the likelihood function to estimate acoustic model. GMMs expose several advantages such as flexibility and capability of representing a large class of sample distributions (Jurafsky and Martin, 2009), insensitiveness to the temporal aspects of the speech (Reynolds, Quantieri and Dunn, 2000).

4.2.1. HMM CONTEXT MODELLING

When building a HMM based speech recognition system, there exist two options: training context-independent (CI) models, known as monophones, or training context-dependent (CD) models. The CI models have the advantage of a good coverage of the training data but they are too general and do not model the contextual phonetic variations (Junqua and Haton, 1996). In contrast, CD model take into consideration the co-articulatory effect of speech is much more consistent (Waibel and Lee, 1990). The most common kind of context dependent model is a triphone HMM (Huang, Acero and Hon, 2001) which represents a phone in left and right context. For example, in the triphone "Z – o + k" that corresponds to SAMPA transcription (Young, 2008) for the Romanian word "joc", the phoneme "o" has as left context "Z" and as right context "k" (Munteanu and Vizitiu, 2008). CD models ensure a better modelling accuracy of speech recognition systems, but the number of triphone models increases heavily when building LVCSR systems. To handle this problem, Young (1992) proposed first model-based tying technique. Another strategy known as state tying is proven to be more efficient (Odell, 1992; Woodland and Odell, 1994) and is generally adopted when modelling triphones. According to this technique, acoustically similar states of the models built for triphones corresponding to each context are grouped together into clusters. The clustering in phonetic classes is achieved using phonetic decision trees.

4.2.2. MODEL ADAPTATION

The AM can be adapted to new speakers in order to achieve an improved accuracy when creating a SI system. In addition to minimize the differences between the model and the new speaker, model adaptation can be used to estimate models on a limited amount of training data. There are three main adaptation techniques, including Maximum A Posteriori (MAP) adaptation, which is the simplest form of acoustic adaptation, Vocal Tract Length Normalization (VTLN), which warps the frequency scale to compensate for vocal tract differences, and Maximum Likelihood Linear Regression (MLLR), which adjusts the Gaussian density parameters or feature vectors so as to increase the likelihood of the adaptation data.

4.3. LEXICAL MODEL

Lexical model is developed to provide the pronunciation of each word from the text scripts corresponding to the audio recordings of speech. This model provides a pronunciation dictionary which contains the list of words and the phone sequence they consist of. It describes how a sequence of sub-units, such as phones, is used to represent larger speech units, such as words from the uttered sentence.

4.4. LANGUAGE MODEL

Language model is the largest component of a LVCSR system and it is developed for detecting the connections between the words in a sentence. This model includes syntactic and semantic rules of a language that constraints the way in which words can be combined into acceptable sentences. Some recognized sentences can be grammatically correct but they are meaningless for Romanian language. For example, the sentence "ceaparoșieestesănătoasă" and the sentence „ceaparoștiieestesănătoasă" are grammatically correct but the second one has no meaning (Cucu, 2014).

If only syntactic constraints are expressed, the LM is reduced to a grammar. Finite state grammars (Aho et al., 1986) are the simplest way of expressing a language model for speech recognition. These grammars are expressed as an unweighted regular expression that represents a finite set of recognizable statements (Zweig, 2004). N-gram models are currently the most widely used LMs in LVCSR. An*n*-gram LM provides the correct word sequence by predicting the likelihood of the *n*-th word on the basis of the *n*-1 preceding words, as it is shown in equation (4):

$$P(W) = P(w_1, w_2, \dots, w_{n-1}, w_n) = \\ = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdot \dots \cdot P(w_n|w_1 w_2 w_3 \dots w_{n-1}) \quad (4)$$

In order to estimate the above probabilities, whether the LM is a unigram (n=1), bigram (n=2) or trigram (n=3) model, it used a method called Maximum Likelihood Estimation (MLE). According to MLE, the parameters of the LM are estimated by taking counts from a corpus and normalizing them so they lie between 0 and 1 (Jurafsky and Martin, 2009).

During the construction of n-gram language models for LVCR systems, two problems are being encountered. Firstly, large amount of training data generally leads to large models which are difficult to handle when building a speech recognition system. In order to overcome this issue, there have been proposed different approaches for reducing the size of LMs, including: count-cutoffs, Weighted Difference pruning, Stolcke pruning, and clustering.

The second problem addresses sparse data faced during the training of domain specific models. This issue is caused by the fact that if a training corpus is limited then some acceptable words sequences of a given language are bound to missing from it. This missing data means that the MLE for some word sequences is zero. Smoothing the probabilities of a language model is essential to deal with the unseen words from the training corpus. In order to deal with data sparsity problem, several smoothing techniques have been developed, such as: additive smoothing, Good-Turing estimate, Jelinek-Mercer smoothing, Katz smoothing, Witten-Bell smoothing, Absolute discounting.

Kneser-Ney smoothing algorithm is the state of the art in this domain (Kneser and Ney, 1995) and it outperforms all other smoothing algorithms in LVCSR applications.

4.5. DECODER

The task of the decoder is to find the best word sequence given the AM and LM. One approach is to search for all possible sequences. However, for LVCSR systems, the search space can become prohibitive. The search space can be described as a finite state machine, where the states are HMMs of words and the transitions between these states are defined by the language model. In recent years, several decoding algorithms have been developed in order to reduce the search space: Stack decoding algorithm, Viterbi algorithm for HMMs (Rabiner, 1989), or multipass search (Huang, Acero and Hon, 2001).

5. ROMANIAN LANGUAGE-CHALLENGES AND DEVELOPED ASR SYSTEMS

Although it is one of the European Union languages, Romanian is still considered a low-resourced language from the point of view of speech and language processing resources. An under-resourced language usually displays some features (Berment, 2014) such as: limited presence on the web, lack of linguistic expertise, lack of electronic resources for natural language processing (NLP) such as monolingual corpora, bilingual electronic dictionaries, and transcribed speech data.

Since 1980's there has been a high interest in developing a speech recognition system for Romanian (Burileanu, L., 1983; Drăgănescu and Burileanu, 1986) and several studies in this field have been driven. Various research workers focused on simple tasks such as vowels recognition (Grigore, Gavăt and Zirra, 1998) or isolated words recognition (Burileanu et al., 1998; Valsan, Sabac and Gavăt, 1998; Sabac, 1998; Burileanu and Popescu, 2004).

Even though there have been made improvements in the field of ASR, the LVCSR problem for Romanian language is far from being solved. The main problem that inhibited the development of large vocabulary tasks is the absence of speech and language resources for Romanian. However, specific speech databases have been created over the years by Romanian research groups but these researches are not widely available.

The latest work in speech recognition is still limited to small-vocabulary tasks. For example, in (Oancea et al., 2004) the authors report a small-vocabulary (approximately 3000 words) continuous speech recognizer in which the number of speakers is limited to 10. The work from (Dumitru and Gavăt, 2008) presents recognition result of 11 speakers for a small-vocabulary task (approximately 4000 words).

In spite of small speech database used for training the speech recognition systems, researchers have successfully adopted different acoustic modelling techniques. For example, Artificial Neural Network (ANN) based approaches are presented in (Dumitru and Gavăt, 2008; Domokoş, 2009), a vector-quantization (VQ) algorithm is illustrated in (Burileanu and Popescu, 2004) and hybrid recognition techniques are proposed in (Dumitru and Gavăt, 2008).

Regarding the lack of Romanian linguistic resources needed for developing a LVCSR system, researchers have tried to overcome this issue using different methodologies. For example, (Cucu et al., 2013) have proposed a methodology that aims to create domain-specific language resources using statistical machine translation.

CONCLUSIONS

In the past decades, several advances in the field of ASR were accomplished for high-resourced language of the world. The availability of speech resources has been facilitating the development of successful ASR commercial solutions. New methods have been created for acoustic and language modeling and the number of deployed speech-based applications reflects the research advances made over the years. Despite the advances, the LVCSR problem is still not solved. There are some variation parameters that influence the high accuracy development of a LVCSR system such as background noise or speaker characteristics.

In Romania, the technology progressed from systems that could recognize digits or a few words to small vocabulary speech recognizers. Due to lack of acoustic and language resources, a LVCSR system has not been developed yet. In order to overcome these issues, several research groups have been focused on creating necessary resources in building such systems, but these resources are not widely available.

ACKNOWLEDGEMENTS

This work was financed by UEFISCI, under PN-II-PTPCCA-2013-4-1644.

References

1. Adami, A.G., 2010. Automaticspeechrecognition: Fromthebeginning to thePortugueselanguage. In *9th International Conferenceon Computacional Processing of thePortugueseLanguage*.
2. Aho, A.V., Sethi, R. and Ullman, J.D., 1986. *Compilers, Principles, Techniques* (pp. 670-671). Addison wesley.
3. O'Shaughnessy, D., 2008. Invitedpaper: Automaticspeechrecognition: History, methods and challenges. *PatternRecognition*, *41*(10), pp.2965-2979.
4. Berment, V., 2004. *Méthodespourinformatiser les langues et les groupes de langues «peudotées»* (Doctoral dissertation, Université Joseph-Fourier-Grenoble I).
5. Baum, L.E., 1972. Anequality and associatedmaximizationtechnique in statisticalestimationforprobabilisticfunctions of Markovprocesses. *Inequalities*, *3*, pp.1-8.
6. Burileanu, C., Teodorescu, V., Stolojanu, G. and Radu, C., 1983. Sistemculogicăprogramatăpentrurecunoaştereacuvintelorizolate, independent de vorbitor. *Inteligenţaartificialăşirobotica, EdituraAcademieiRomâne, Bucureşti*, *1*, pp.266-274.
7. Burileanu, D., Sima, M., Burileanu, C. and Croitoru, V., 1998. A Neural Network-Based Speaker-IndependentSystemfor Word Recognition in RomanianLanguage. In *Proceedings of TheFirst Workshop on Text, Speech and Dialogue, Brno, CzechRepublic* (pp. 177-182).
8. Burileanu, C. and Popescu, V., 2004. Anefficientdistributedspeechrecognitionfrontendimplementationusing a Motorola Star Core 140 basedplatform. *PolitehnicaUniversity of TimisoaraScientificBulletin-TransactionsonElectronics and Communications*, *49*(63-1), pp.305-310.

9. Cucu, H., Buzo, A., Besacier, L. and Burileanu, C., 2014. SMT-based ASR domainadaptationmethodsforunder-resourcedlanguages: Application to Romanian. *SpeechCommunication*, *56*, pp.195-212.
10. Davis, S. and Mermelstein, P., 1980. Comparison of parametricrepresentationsformonosyllabicwordrecognition in continuouslyspokensentences. *IEEE transactionsonacoustics, speech, and signalprocessing*, *28*(4), pp.357-366.
11. Davies, K., Donovan, R.E., Epstein, M., Franz, M., Ittycheriah, A., Jan, E.E., LeRoux, J.M., Lubensky, D., Neti, C., Padmanabhan, M. and Papineni, K., 1999, September. The IBM conversationaltelephonysystemforfinancialapplications. In *Eurospeech*.
12. Drăgănescu, M. and Burileanu, C., 1986. Analiza şisintezasemnalului vocal. *EdituraAcademiei*.
13. Domokoş, J., 2009. *Contributionsoncontinuousspeechrecognition and natural languageprocessing* (Doctoral dissertation, PhD Thesis, TechnicalUniversity of Cluj-Napoca, Cluj-Napoca, Romania).
14. Dumitru, C.O. and Gavat, I., 2008. *Progress in speechrecognitionforRomanianlanguage*. INTECH Open Access Publisher.
15. Gao, Y., Ramabhadran, B., Chen, J., Erdogan, H. and Picheny, M., 2001. Innovativeapproachesforlargevocabularynamerecognition. In *Acoustics, Speech, and SignalProcessing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conferenceon* (Vol. 1, pp. 53-56). IEEE.
16. Grigore, O., Gavat, I. and Zirra, M., 1998. Neural networkvowelrecognition in Romanianlanguage. *Proceedings CONTI'98*, pp.165-172.
17. Ghai, W. and Singh, N., 2012. Literaturereviewonautomaticspeechrecognition. *International Journal of ComputerApplications*, *41*(8).
18. Gruhn, R.E., Minker, W. and Nakamura, S., 2011. *Statisticalpronunciationmodelingfor non-nativespeechprocessing*. SpringerScience& Business Media.
19. Hermansky, H., 1989. Perceptual linear predictive (PLP) analysis of speech. *theJournal of theAcousticalSociety of America*, *87*(4), pp.1738-1752.
20. Huang, X., Acero, A., Hon, H.W. and ForewordBy-Reddy, R., 2001. *Spokenlanguageprocessing: A guide to theory, algorithm, and systemdevelopment*. Prentice hall PTR.
21. Imseng, D., Motlicek, P., Bourlard, H. and Garner, P.N., 2014. Usingout-of-language data to improveanunder-resourcedspeechrecognizer. *Speechcommunication*, *56*, pp.142-151.
22. Jan, E.E., Maison, B., Mangu, L. and Zweig, G., 2003, September. Automaticconstruction of uniquesignatures and confusable sets for natural languagedirectoryassistanceapplications. In *INTERSPEECH*.
23. Junqua, J.C. and Haton, J.P., 2012. *Robustness in automaticspeechrecognition: Fundamentals and applications* (Vol. 341). SpringerScience& Business Media.
24. Jurafsky, D. and Martin, J.H., 2009. *Speech and languageprocessing*. Pearson.
25. Kneser, R. and Ney, H., 1995, May. Improvedbacking-off for m-gramlanguagemodeling. In *Acoustics, Speech, and SignalProcessing, 1995. ICASSP-95., 1995 International Conferenceon* (Vol. 1, pp. 181-184). IEEE.
26. Munteanu, D.P. and Vizitiu, C.I., 2008. RobustRomanianlanguageautomaticspeechrecognizerbasedonmultistyle training. *WSEAS TransactionsonComputerResearch*, *3*(2), pp.98-109.
27. Oancea, E., Gavăt, I., Dumitru, O. and Munteanu, D., 2004. ContinuousspeechrecognitionforRomanianlanguagebasedoncontextdependentmodeling. *Communications*, pp.221-224.
28. Odell, J.J., 1992. The use of decisiontreeswithcontextsensitivephonememodelling. *Master'sthesis, Cambridge University, Cambridge, England*.
29. Ordowski, M., Deshmukh, N., Ganapathiraju, A., Hamaker, J. and Picone, J., 1999. A publicdomainspeech-to-textsystem. In *EUROSPEECH*.
30. Pellom, B., Ward, W., Hansen, J., Cole, R., Hacioglu, K., Zhang, J., Yu, X. and Pradhan, S., 2001, March. University of Colorado dialogsystemsfortravel and navigation. In *Proceedings of thefirstinternationalconferenceon Human languagetechnologyresearch* (pp. 1-6). AssociationforComputationalLinguistics.

31. Rabiner, L., Levinson, S., Rosenberg, A. and Wilpon, J.A.Y.G., 1979. Speaker-independentrecognition of isolatedwordsusingclusteringtechniques. *IEEE TransactionsonAcoustics, Speech, and SignalProcessing*, *27*(4), pp.336-349.Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., 2000. Speaker verificationusingadaptedGaussian mixture models. *Digital signalprocessing*, *10*(1), pp.19-41.
32. Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., 2000. Speaker verificationusingadaptedGaussian mixture models. *Digital signalprocessing*, *10*(1), pp.19-41.
33. Saon, G., Padmanabhan, M., Gopinath, R. and Chen, S., 2000. Maximumlikelihooddiscriminantfeaturespaces. In *Acoustics, Speech, and SignalProcessing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conferenceon* (Vol. 2, pp. II1129-II1132). IEEE.
34. Tiwari, V., 2010. MFCC and itsapplications in speaker recognition. *International JournalonEmerging Technologies*, *1*(1), pp.19-22.
35. Trandabăț, D., Irimia, E., Mititelu, V.B., Cristea, D. and Tufiș, D., 2012. Limbaromânăîn era digitală–TheRomanianLanguage in the Digital Age. META-NET White Paper Series: Europe'sLanguages in the Digital Age.
36. Valsan, Z., Sabac, B., Gavat, I. and Zamfirescu, D., 1998. Combiningself-organizingmap and multilayerperceptron in a neural systemforimprovedisolatedwordrecognition. *ProceedingsCommunications' 98*, pp.245-251
37. Waibel, A., 1990. *Readings in speechrecognition*. Morgan Kaufmann.
38. Wegmann, S., McAllaster, D., Orloff, J. and Peskin, B., 1996, May. Speaker normalizationonconversationaltelephonespeech. *IEEE International ConferenceonAcoustics, Speech, and SignalProcessing*.Vol. 1, p. 339-341.
39. Woodland, P., Young, S.: The HTK Tied-State Continuous Speech Recognizer. EUROSPEECH. ESCA, Berlin, Germany (1993) 2207-2210.
40. Woodland, P.C., Odell, J.J., Valtchev, V. Young, S.J., 1994. Largevocabularycontinuousspeechrecognitionusing HTK. In *Acoustics, Speech, and SignalProcessing, 1994. ICASSP-94., IEEE International Conferenceon* ,Vol. 2, p. II-125.
41. Young, S.J., 1992. The general use of tying in phoneme-based HMM speechrecognisers. In *Acoustics, Speech, and SignalProcessing, 1992. ICASSP-92., IEEE International Conferenceon* Vol. 1, p. 569-572
42. Young, S., 2008. HMMs and relatedspeechrecognitiontechnologies. In *SpringerHandbook of SpeechProcessing* (pp. 539-558). SpringerBerlin Heidelberg.
43. Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J. and Hetherington, L., 2000. JUPlTER: a telephone-basedconversational interface forweatherinformation. *IEEE Transactionsonspeech and audio processing*, *8*(1), pp.85-96.
44. Zweig, G. and Picheny, M., 2004. Advances in largevocabularycontinuousspeechrecognition. *Advances in Computers*, *60*, pp.249-291.