On transcribing informally-pronounced numbers in Romanian speech

Horia Cucu, Alexandru Caranica, Andi Buzo, and Corneliu Burileanu

***Abstract*—The pronunciation model, a mapping between the lexicon words and their phonetic representation, has a key role in automatic speech recognition. Although many times neglected, the accuracy of this model influences significantly the accuracy of the whole system. This study discusses within-word and cross-word pronunciation variations for Romanian numbers and proposes the solutions to model them in the phonetic dictionary and the language model of an existing speech recognition system for Romanian. The evaluation is performed of a read speech corpus comprising rational numbers with up to three decimal digits. The experiments show a relative WER improvement of 14% over the baseline when within-word pronunciation variations are taken into account and an additional relative WER improvement of 63% when cross-word pronunciation variations are also modelled.**

***Keywords*—automatic speech recognition, phonetic dictionary, pronunciation modelling, cross-word pronunciation variation, within-word pronunciation variation**

I. Introduction

THE various sources of speech variability, such as foreign or regional accents, speaker physiology, speaking style (read or conversational), speaking rate, emotional speech, etc., still represent real challenges for Automatic Speech Recognition (ASR) systems [1]. All the above lead to a high degree of variability in the pronunciation of words. Pronunciation variation does not refer only to the simple coarticulation of phones in a specific context. It is also manifested in the form of insertions, deletions, or substitutions of phonemes relative to the canonical transcription of the dictionary words ("what's" pronounced /waz/ instead of /wats/). If these pronunciation variations are not modelled in a form or another in the ASR system, then its accuracy is significantly influenced when it comes to transcribing conversational or accented speech.

Depending on how the pronunciation variations span over one or multiple adjacent words, they are classified as within-word or cross-word variations. Within-word variations manifest as alternative pronunciations of single words and are usually addressed by adding pronunciation variants to the phonetic dictionary. Cross-word pronunciation variations appear due to cliticization, contraction and reduction effects [2] at the boundary between adjacent words and result in alterations of the last phones of the first word and the first few phones of the second word. In this case, the altered pronunciations cannot be added to phonetic dictionary because they are valid only in certain word contexts, while in most others they increase the confusability of words.

Manuscript received February 5, 2015. This work was supported in part by the Sectoral Operational Programme "Human Resources Development" 2007-2013 of the Ministry of European Funds through the Financial Agreements POSDRU /159/1.5/S/132395 and POSDRU /159/1.5/S/134398 and in part by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, throught project no. 332/2014.

H. Cucu, A. Caranica, A. Buzo and C. Burileanu are with the Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania (e-mail: {horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro).

Both within-word and cross-word pronunciation variations can be approached in a knowledge-based fashion or in a data-driven fashion. In the first case, linguistic knowledge is required to develop a set of phonological rules and apply them systematically to the words in the baseline pronunciation dictionary, generating alternative pronunciations. For example, in [2] the authors apply rules, such as /n/-deletion, /r/-deletion, /t/-deletion, /ə/-deletion and /ə/-insertion, for improving a Dutch ASR system. AbuZeina et al. apply three other phonological rules for Arabic: /n/ -> /m/, /n/ -> /aɪ/ and /t/ -> /d/ [3]. This approach was also used with various rules for specific English words/phrases (sort of -> sorta, got you -> gotcha, going to -> gonna, etc.) in [4][5][6] and also for common French phonological phenomena, such as liaisons and mute-e elisions [7].

Data-driven approaches to pronunciation variation modelling imply using a phone recognizer to obtain alternative pronunciations of words and word sequences, select the most frequent ones and add them to the phonetic dictionary. Data-driven approaches were proposed in [6] and [8].

Regardless of how the alternative pronunciations are obtained, within-word pronunciation modelling is much simpler to implement than cross-word pronunciation modelling. Because the within-word alternative pronunciations refer strictly to single words, they can be simply added to the phonetic dictionary. Most of the speech recognition decoders are able to deal with phonetic dictionaries comprising multiple pronunciations for the words. Cross-word alternative pronunciations cannot be added to the phonetic dictionary because they represent pronunciations for word sequences, not single words. The general methodology used for modelling these pronunciations is to design artificial compound-words (also called multi-words [5] or phrases [9]) and to provide special pronunciations for these new tokens [8]. Knessens et al. also tried a different approach: adding directly the alternative pronunciations of every composing word in a compound-word to the phonetic dictionary, but concluded that this method does not work as good as the one discussed above [2].

In this context, this study deals with an important and very frequent pronunciation variation in the Romanian language: informal pronunciation of numbers. Some numbers are written as single words, but most of them are composed of several words, so both within-word and cross-word pronunciation solutions are needed. We approach the issue in a knowledge-based fashion by identifying the explicit phonological rules

which are used to pronounce informally the Romanian numbers. For cross-word variations we design artificial compound-words and provide special pronunciations for them in the phonetic dictionary. The experimental results section reports significant relative improvements of the speech recognition word error rates (WERs) after the implementation of the proposed solutions. To the best of our knowledge, this is the first study to deal cross-word pronunciation modelling for the Romanian language.

The rest of the paper is organized as follows. Section II presents in detail the problem and the proposed solution, Section III discusses the experimental setup and the results, while the last part is used to draw the conclusion.

II. Informal Pronunciation of Romanian Numbers

In the Romanian language the numbers representing amounts of money, dates, etc. are heavily pronounced informally. Although in read, prepared speech numbers are usually pronounced correctly, most of their occurrences in free or spontaneous speech are informal pronunciations. This is a real challenge for ASR systems which use a phonetic dictionary comprising canonical pronunciations for the vocabulary words. This section describes the phenomena observed in pronouncing informally the Romanian numbers and proposes solutions for within-word and cross-word variations.

After a thorough study of Romanian numbers pronunciations we came to the conclusion that informal pronunciations occur more often in two digit numbers. Consequently, these numbers will be treated in more detail in the following subsection.

A. Romanian Two-Digit Numbers

Two-digit Romanian numbers are formed similarly to two-digit English numbers. There are two separate rules for the groups 10 - 19 and 20 - 99.

The numbers between 10 and 19 are written as compound-words formed by concatenating the unit words: 1, 2, ..., 9 (”un” / ”unu”, ”doi”, ..., ”nouă” in Romanian), with the preposition ”to” (”spre” in Romanian) and with the word ”ten” (”zece” in Romanian). These numbers are summarized in Table I. There are two exceptions (14 and 16) for which the unit word is slightly modified ”pai” instead of ”patru” and ”şai” instead of ”şase” (similarly to the English exception ”fif” instead of ”five” in ”fifteen”).

The numbers between 20 and 99 are written as word phrases (separate words) by joining the compound word for tens: 20, 30, ..., 90 (”douăzeci”, ”treizeci”, ..., ”nouăzeci” in Romanian) with the conjunction ”and” (”şi” in Romanian) and with the unit words: 1, 2, ...,9. A part of these numbers (30 - 39) are summarized in Table II. There are no exceptions to this composition rule.

It is worth mentioning that, similarly to English, Romanian numbers of any size are formed based on two-digit numbers and special words denoting hundreds, thousands, etc. (e.g. ”*thirty two* thousand six hundred *fifty seven*” is ”*treizeci şi două* de mii şase sute *cincizeci şi şapte*” in Romanian).

TABLE I: Two-digit Romanian Numbers (10 - 19)

Number	Text Version (English)	Text Version (Romanian)
10	ten	zece
11	eleven	unsprezece
12	twelve	doisprezece
13	thirteen	treisprezece
14	fourteen	paisprezece
15	fifteen	cincisprezece
16	sixteen	şaisprezece
17	seventeen	şaptesprezece
18	eighteen	optsprezece
19	nineteen	nouăsprezece

TABLE II: Two-digit Romanian Numbers (30 - 39)

Two-digit Number	Text Version (English)	Text Version (Romanian)
30	thirty	treizeci
31	thirty-one	treizeci şi unu
32	thirty-two	treizeci şi doi
33	thirty-three	treizeci şi trei
34	thirty-four	treizeci şi patru
35	thirty-five	treizeci şi cinci
36	thirty-six	treizeci şi şase
37	thirty-seven	treizeci şi şapte
38	thirty-eight	treizeci şi opt
39	thirty-nine	treizeci şi nouă

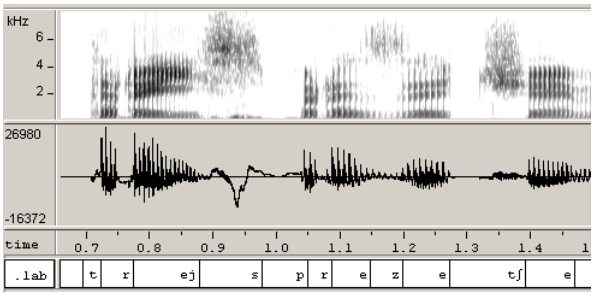
B. Within-word and Cross-word Pronunciation Variation

The two-digit numbers in the first group mentioned above (10 - 19) are usually pronounced informally by changing several syllables (within the compound word) into a shorter one. For example, the word ”trei.spre.ze.ce” (13), formally pronounced /trej.spre.ze.ʧe/, is usually pronounced /trej.ʃpe/. As you can notice from this example, the syllables ”spre”, ”ze” and ”ʧe” have been merged and changed into ”şpe”. Figure 1 illustrates this behaviour, showing both the correct, formal pronunciation and the informal pronunciation of the number 13. This pronunciation variation can be seamlessly integrated in the pronunciation model by adding a second pronunciation for all these words.

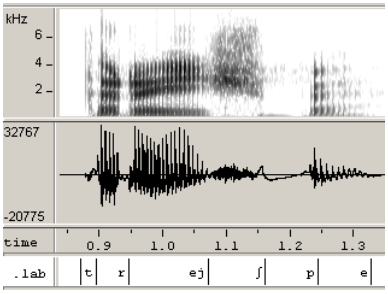
Most of the two-digit numbers in the second group mentioned above (21 - 29, 31 - 39, ... 91 - 99) are written as three consecutive words (e.g. the number 36 is written ”treizeci şi şase”). They are usually pronounced informally by reducing one or several syllables. There are two commonly used informal pronunciations for these numbers. The phrase ”trei.zeci şi şa.se” (36), formally pronounced /trej.zeʧ ʃi ʃa.se/, is usually pronounced /trej.ze ʃi ʃa.se/ (the /ʧ/ in the second syllable is missing) or /trej.ʃa.se/ (the second syllable and the second word are missing). Figure 2 illustrates this behaviour, showing both the correct, formal pronunciation and the informal pronunciations of the number 36. As exemplified, this pronunciation variation spreads across several words and cannot be integrated seamlessly in the pronunciation model, because this model stores words (not word sequences) pronunciations.

C. Solution to Cross-word Pronunciation Variation

This section proposes a method for modelling the cross-word pronunciation variations described above in an automatic



(a) Canonical Pronunciation of the Number 13



(b) Informal Pronunciation of the Number 13

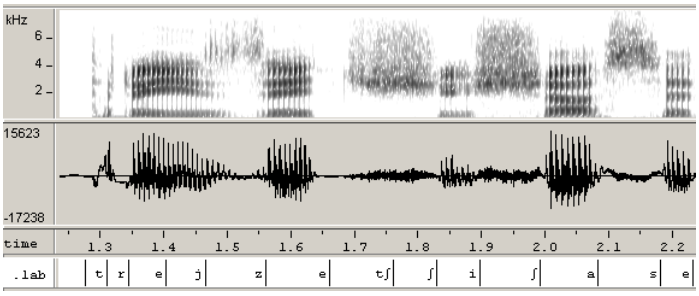
Fig. 1: Pronunciations of the Number 13 (”treisprezece” in Romanian)

speech recognition system. To solve this problem we propose the following steps:

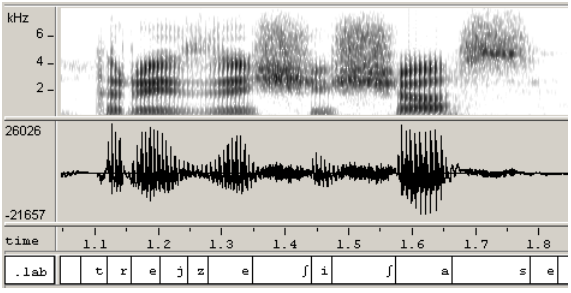
1) identify the phrases for which cross-word pronunciation variation occurs (some two-digit numbers),
2) merge the sequence words for which the variation was observed into an artificial compound-word,
3) specify the various pronunciations for these artificial compound-words in the pronunciation model,
4) update the language model to use these compound-words instead of original sequences of words.

The second step in the algorithm is performed by merging several words using an underscore (e.g. ”treizeci şi şase” -> ”treizeci_şi_şase”). In this case, the ASR system will output (after decoding) regular Romanian words and artificial compound-words such as the one in the previous example. In an output post-processing stage, all underscore characters can be replaced with space so that the final transcription contains only regular Romanian words. Note that in Romanian there are many regular compound-words, which are connected using a hyphen (-) and this is the reason why the hyphen character cannot be used in our algorithm.

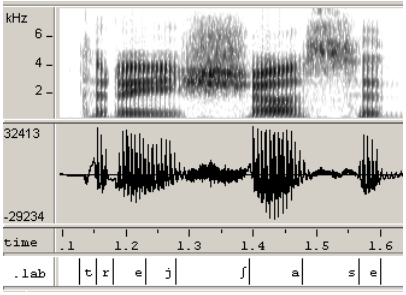
The forth step in the algorithm (the modification of the language model) depends on whether the target ASR task is a continuous speech task or a rule-grammar task. This paper deals with the second case. Rule grammars model explicitly all speech recognition candidates: all the word sequences which can be uttered by the speaker (along with their probabilities). In this case, the modification of the language model means replacing grammar rules for word sequences with rules for artificial compound-words. An example of grammar modification is presented in Figures 3 and 4. The figures present rule grammars able to recognize Romanian integer numbers with up to three digits. Figure 3 shows the original rule grammar in



(a) Canonical Pronunciation of the Number 36



(b) Informal Pronunciation #1 of the Number 36



(c) Informal Pronunciation #2 of the Number 36

Fig. 2: Pronunciations of the Number 36 (”treizeci şi şase” in Romanian)

which two digit numbers between 21 and 99 were modelled as sequences of words. Figure 4 illustrates the modified grammar in which the sequences of words representing two-digit numbers were merged into artificial compound-words so that cross-word pronunciation variation can be taken into account.

III. Experimental Setup and Results

A. Baseline ASR system

All the experiments presented onwards were made using the speech recognition system for the Romanian language developed by the Speech and Dialogue Research Laboratory [10]. The ASR system is built upon the CMU Sphinx speech recognition toolkit [11]. More specifically, the decoding system uses the CMU Sphinx 4 Java decoder.

The acoustic models are speaker-independent, 3-state HMMs with output probabilities modelled with GMMs. The classic Mel Frequency Cepstral Coefficients (MFCCs) plus their first and second temporal derivatives (13 MFCCs + deltas + double deltas) were used as speech features. The 36 phonemes of the Romanian language were modelled contextually (context dependent phonemes) with 4000 HMM senones.

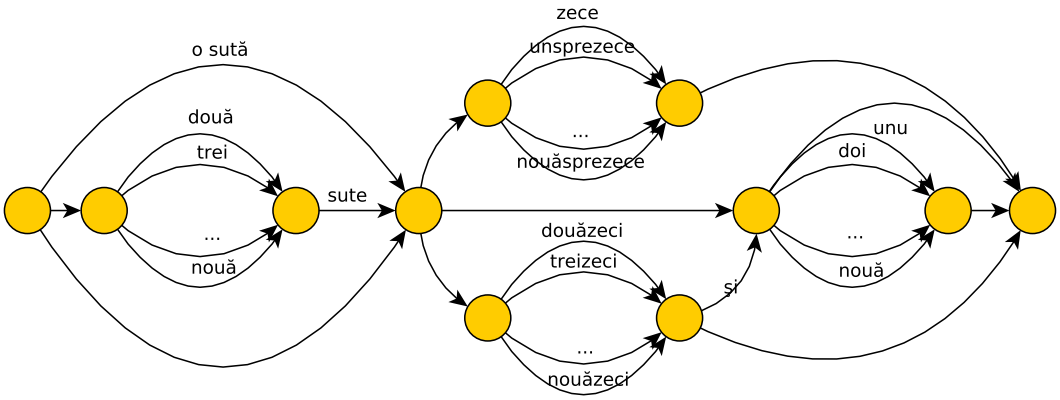


Fig. 3: Original Romanian Numbers Grammar

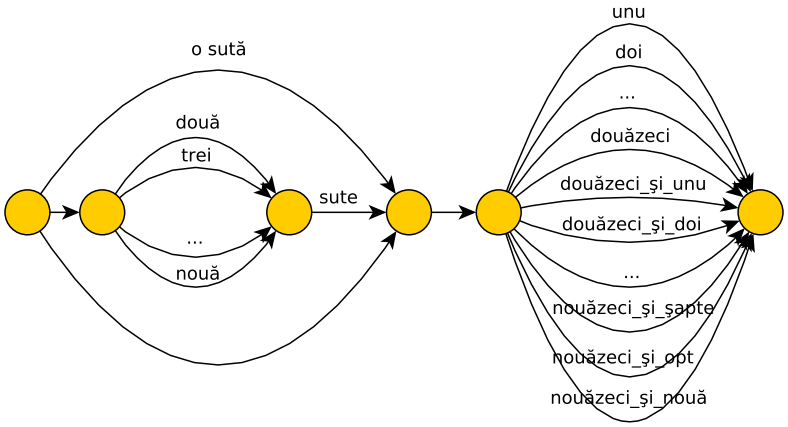


Fig. 4: Modified Romanian Numbers Grammar

The number of Gaussian mixtures per senone state is variable (32/64/128), adapted to the size and variability of the training speech corpus (in the following experiments we used 32 GMMs per senone). The acoustic models were trained and optimized on a Romanian read and spontaneous speech corpus and using the CMU Sphinx Toolkit.

TABLE III: Evaluation Speech Corpus

Pronunciation	Utterances	Words	Speakers
Formal	1150	4949	12
Informal	1850	7723	13
Total	3000	12672	14

B. Evaluation Speech Corpus

The evaluation speech corpus used in the experiments is part of a larger evaluation Romanian speech corpus comprising task-specific utterances for the following speech recognition tasks: numbers, dates, cities, forenames, surnames and yes/no.

The corpus was recently created by recording various pre-defined phrases representing in-grammar utterances for the six ASR tasks mentioned above. The phrases were chosen randomly with the goal of covering as much as possible these speech recognition tasks. The recordings were made using an online recording application previously developed by the our research group. Fourteen speakers were involved voluntarily in the speech corpus development process. Some of them recorded all the phrases for all ASR tasks, while others recorded only partially these phrases.

For the purpose of evaluating the recognition of informally pronounced numbers, the utterances for the Numbers ASR task were recorded in a special manner. The speakers were asked to pronounce informally the first 150 utterances and formally the other 100 utterances. Consequently, a part of the Numbers speech corpus (labelled ”inf”) comprises informally pronounced numbers, while the other part (labelled ”for”) comprises formally pronounced numbers. All these utterances comprise rational numbers with up to three decimal digits between minus one billion and plus one billion. The details of the Numbers speech corpus are provided in Table III.

C. Experimental Results

Several ASR setups were evaluated on the speech corpus described above. In all cases, the same acoustic model and speech decoder (described in Section III-A) were used. The pronunciation models and rule grammars differ from one

TABLE IV: The effects of informal pronunciations of Romanian numbers

Pronunciation Model	Rule Grammar	WER[%] inf	for	all
Formal pronunciations	Regular numbers grammar	26.2	2.4	16.9
+ within-word informal pronunciations	Regular numbers grammar	22.5	1.8	14.5
+ cross-word informal pronunciations	+ 21-99 modelled as compound words	8.2	2.2	6.0

experimental setup to another as follows. The baseline system uses (i) a pronunciation model comprising only formal pronunciations of Romanian numbers and (ii) a regular grammar for recognizing rational numbers with up to three decimal places between minus one billion and plus one billion.

The first system enhancement involved providing within-word informal pronunciations in the pronunciation dictionary. In this experimental setup the same regular rule grammar is used (there are no other new words to be modelled).

The second system enhancement involved (i) switching to a modified rule grammar in which the two-digit numbers are modelled as artificial compound-words and (ii) providing cross-word informal pronunciations for these compound-words in the pronunciation dictionary. The differences between the regular rule grammar and the modified grammar are exemplified in Figures 3 and 4, which present the differences for a simpler rule grammar.

The experimental results on the Numbers speech corpus and its two parts (”inf” and ”for”) are presented in Table IV. In the first setup informal pronunciations are not modelled at all (neither within-word, nor cross-word variations). In the second setup within-word informal pronunciation variations are inserted into the pronunciation model for the two-digit numbers between 10 and 19 (these numbers are written with single words, see Table I). Finally, in the third setup, the rule grammar is modified to model two-digit numbers between 20 and 99 as artificial compound-words (these numbers are normally written with several words, see Table II) and cross-word informal pronunciation variations (for these compound-words) are inserted into the pronunciation model.

Table IV shows that the pronunciation modelling approaches proposed in this paper have significant beneficial effects for the task of recognizing informally pronounced numbers. Within-word pronunciation modelling of informal numbers brings a relative WER improvement of 14% over the baseline. Furthermore, cross-word pronunciation modelling of informal numbers brings a relative WER improvement of 63% over the previous pronunciation modelling technique. As expected, on formally pronounced numbers the results are more or less the same regardless of the pronunciation modelling technique. However, it is interesting to see that even in the third setup the WER on informally pronounced numbers is much higher than the WER on formally pronounced numbers. This means that there is still room for improving the recognition of informally pronounced numbers.

IV. Conclusion

This paper presented the pronunciation modelling techniques used to approach the informal pronunciation of Romanian numbers in the context of automatic speech recognition. Both within-word and cross-word techniques were employed, because informally pronounced Romanian numbers between 10 and 19 are written as single words, while numbers between 20 and 99 are written as sequences of words. Only the pronunciation of two-digit numbers were treated because, similarly to English, Romanian numbers of any size are formed based on them (e.g. ”*thirty two* thousand six hundred *fifty seven*” is ”*treizeci şi două* de mii şase sute *cincizeci şi şapte*” in Romanian).

The pronunciation modelling techniques were evaluated on a read speech corpus comprising rational numbers with up to three decimal digits between minus one billion and plus one billion. The corpus was explicitly recorded for the experiments presented in the study and contains both formal and informal pronunciations recorded by 14 speakers. The experiments showed a relative WER improvement of 14% over the baseline when within-word pronunciation variations are taken into account and an additional relative WER improvement of 63% when cross-word pronunciation variations are also modelled.

In the near future we plan to address informal pronunciations in context, in continuous speech. Applying the cross-word pronunciation technique for continuous speech is more difficult because the artificially-created compound words (e.g. treizeci_şi_două) cannot be easily modelled in n-gram language models. Several solutions were envisioned and we are currently running experiments on this subject.

References

[1] M. Benzeguiba, R. De Mori et al., “Automatic speech recognition and speech variability: a review,” *Speech Commun*, vol. 49, no. 10-11, pp. 763–786, Nov. 2007.

[2] J. M. Kessens, M. Wester, H. Strik, “Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation,” *Speech Commun*, vol. 29, no. 2-4, pp. 193–207.

[3] D. AbuZeina, W. Al-Khatib, M. Elshafei, H. Al-Muhtaseb, “Cross-word Arabic pronunciation variation modeling for speech recognition,” *International Journal of Speech Technology*, vol. 14, no. 3, pp. 227–236.

[4] E. P. Giachin, A. E. Rosenberg, C.-H. Lee, “Word juncture modeling using phonological rules for HMM-based continuous speech recognition,” *Comput Speech Lang*, vol. 5, no. 2, pp. 155–168.

[5] M. Finke, A. Waibel, “Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition,” in *Proc. Int. Conf. Eurospeech*, Rhodes, Greece, 1997, pp. 2379–2382.

[6] M. Wester, “Pronunciation modeling for ASR - knowledge-based and data-derived methods,” *Comput Speech Lang*, vol. 17, no. 1, pp. 69–85.

[7] G. Perennou, L. Pousse, “Dealing with pronunciation variants at the language model level for automatic continuous speech recognition of French,” in *Proc. Int. Conf. Eurospeech*, Rhodes, Greece, 1997, pp. 2727–2730.

[8] G. Saon, M. Padmanabhan, “Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition,” *Trans. Speech and Audio Proc.*, vol. 9, no. 4, pp. 327–332, May 2001.

[9] K. Ries, F. D. Buo, A. Waibel, “Class phrase models for language modeling,” in *Proc. Int. Conf. Speech Language Processing*, Philadelphia, USA, 1996, pp. 398–401.

[10] H. Cucu, A. Buzo, L. Petrică, D. Burileanu and C. Burileanu, “Recent Improvements of the SpeeD Romanian LVCSR System,” in *Proc. Int. Conf. Communications (COMM)*, Bucharest, 2014, pp. 111–114.

[11] P. Lamere, P. Kwok, W. Walker, E. Gouvea, R. Singh, B. Raj, and P. Wolf, “Design of the CMU Sphinx-4 decoder,” in *Proc. Int. Conf. Interspeech*, Geneve, Switzerland, 2003, pp. 1181–1184.