ASR errors in transcribing informal pronunciations of Romanian numbers

Horia Cucu, Andi Buzo, Corneliu Burileanu

Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania

{horia.cucu, andi.buzo, corneliu.burileanu}@upb.ro

Abstract

In Romanian spontaneous speech, numbers are almost always pronounced informally or colloquially. Automatically transcribing colloquially-pronounced numbers tends to be very challenging due to the fact that the pronunciation variants are very different from the canonical pronunciations. Within-word and especially cross-word pronunciation variations (contractions and elisions) lead to numerous ASR errors. The problem is considered critical in many applications because numbers usually carry important information (such as amounts of money, dates, times, etc.) which cannot be inferred from the context. This paper analyses in-depth these errors and the underlying phonological phenomena and proposes a method to prevent their occurrence. The effectiveness of the proposed solution is evaluated on a read speech corpus comprising Romanian numbers. A relative word error rate (WER) improvement of 12% over the baseline is obtained when within-word pronunciation variations are taken into account and an additional relative WER improvement of 35% is obtained when cross-word pronunciation variations are also modelled.

Index Terms: speech recognition, pronunciation modelling, cross-word pronunciation variation, language modelling, error analysis

1. Introduction

In the last decade the speech recognition research community has focused a lot on acoustic and language modelling. New and efficient acoustic and language modelling techniques have been developed and this proved to bring important improvements to state-of-the-art automatic speech recognition (ASR). However, this trend has also triggered a drop in attention for pronunciation modelling. The pronunciation dictionary, the third main component of an ASR system, links the lexicon words with corresponding pronunciation variants. Especially when creating an ASR system for a new language, the development of the pronunciation (or phonetic) dictionary is of critical importance in order to ensure acceptable ASR performance [1]. However, the first version of the pronunciation dictionary usually comprises only canonical pronunciations for the words. Creating a good pronunciation dictionary involves predicting pronunciation variations, that is, alternative pronunciations observed for a linguistically identical word. This is a complicated problem and depends on many factors, such as the linguistic origin of the speaker, the speaker's education and socio-economic level, the speaking style and conversational context and the relationship between speakers [2].

Pronunciation variability modelling has been widely addressed in the past. The various methods that have been proposed to generate pronunciation variants can be grouped in two categories: knowledge-based approaches and data-driven approaches. 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. The rules may produce "parallel variants" (i.e. variants in which some phonemes may be replaced with others) or "sequential variants" (i.e. variants which allow some phonemes to be optional) [3]. For example, in [4] the authors apply phonological rules, such as /n/-deletion, /r/-deletion, /t/-deletion, /ə/-deletion and /ə/-insertion, creating sequential pronunciation variants with the goal of improving a Dutch ASR system. AbuZeina et al. apply three other phonological rules (/n/ -> /m/, /n/ -> /aɪ/ and /t/ -> /d/), generating parallel pronunciation variants for an Arabic ASR system. [5]. This approach was also used for common French phonological phenomena, such as liaisons and schwa elisions [6].

One drawback of knowledge-based methods is that they are not language-independent and do not always capture the irregularities in natural languages. Therefore, for the cases where a draft pronunciation dictionary is already available, alternative approaches have been proposed to introduce pronunciation variants. These data-driven methods are based on the idea that, given enough examples, it should be possible to predict the pronunciation of unseen words or generate pronunciation variants for known words. Various data-driven methods were proposed in the literature so far: starting from neural networks [7], confusion tables [8], to statistical machine translation [9] and automatic generation of rules for phoneme-to-phoneme conversion [10].

With the availability of large corpora of transcribed speech corpora, another type of data-driven approaches emerged. These involve using a phoneme recognizer to over-generate pronunciation variants for the words in a large speech corpus, then selecting the most frequent ones and finally adding them to the phonetic dictionary. Such data-driven approaches were proposed in [11] and [12]. This type of data-driven methods were also used to study more thoroughly the phenomena of liaisons and optional schwa observed in French in order to derive rules that can be used to generate pronunciation variants [3][13].

All the studies mentioned above have focused on generating pronunciation variants for single words. However, there are cases in which the pronunciation variations span across word boundaries and cases when specific pronunciation variants occur only in a certain context. These cross-word pronunciation variations appear due to cliticization, contraction and reduction effects [4] 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. The pronunciation variants for the composing words cannot be simply added to the phonetic dictionary because they are valid only in certain word contexts, while in the rest they are potential sources of errors (as they may increase the homophones rate). Specific cross-word pronunciation variations (sort of -> sorta, got you -> gotcha, going to -> gonna, etc.) were dealt with in [11][14][15]. The general

Table 1: Examples of two-digit Romanian numbers

No.	English	Romanian	Canonical Pron.	Informal Pron. #1	Informal Pron. #2
10	ten	zece	/ze.ʧe/	-	-
11	eleven	unsprezece	/un.spre.ze.ʧe/	/un.ʃpe/	-
12	twelve	doisprezece	/doj.spre.ze.ʧe/	/doj.ʃpe/	-
13	thirteen	treisprezece	/trej.spre.ze.ʧe/	/trej.ʃpe/	-
19	nineteen	nouăsprezece	/no.wə.spre.ze.ʧe/	/no.wə.ʃpe/	-
30	thirty	treizeci	/trej.zeʧʲ/	-	-
31	thirty-one	treizeci şi unu	/trej.zeʧʲ ʃi unu/	/trej.ze ʃi unu/	/trej.ʃunu/
37	thirty-seven	treizeci şi şapte	/trej.zeʧʲ ʃi ʃap.te/	/trej.ze ʃi ʃap.te/	/trej.ʃap.te/
38	thirty-eight	treizeci şi opt	/trej.zeʧʲ ʃi opt/	/trej.ze ʃi opt/	/trej.ʃopt/
39	thirty-nine	treizeci şi nouă	/trej.zeʧʲ ʃi no.wə/	/trej.ze ʃi no.wə/	/trej.ʃno.wə/

methodology used for modelling these pronunciations is to design artificial compound-words (also called multi-words [14] or phrases [16]) and to provide special pronunciations for these new tokens [12].

In this paper we continue the work started in [17], where we proposed a solution to deal with an important and very frequent pronunciation variation in the Romanian language: informal pronunciation of numbers. In most cases, especially in spontaneous speech, Romanian numbers are uttered colloquially, using pronunciation variants which are very different from the canonical pronunciations. The phonetic variations span across several syllables within words and also across words, similarly to the English "got you -> gotcha" variation. While in the afore mentioned work we solved this problem for rule-grammar ASR systems, in this study we propose a method to deal with these variations in large vocabulary ASR systems and explore in-depth the various error types influencing the accuracy of the recognition process. Because we are dealing with specific pronunciation variants affecting only words representing numbers, we proposed a knowledge-based method of modelling these variations.

The rest of the paper is organized as follows. In section 2 we describe the problem in detail, giving examples of ASR errors and propose a solution to address them. Section 3 presents the experimental setup and the ASR results. In section 4 we analyse the errors to better understand what worked and what did not work. The last section is dedicated to the conclusions.

2. Informal pronunciation of numbers in the Romanian language

In Romanian spontaneous speech most of the numbers are pronounced colloquially (different from the literary speech pronunciation), while strict pronunciation rules are respected only in formal speeches. This section describes the phonological phenomena observed in Romanian number pronunciations and proposes modelling solutions for pronunciation variations in the context of automatic speech recognition.

Similarly to English, Romanian numbers with more than three digits are formed based on two-digit numbers and other 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). After a thorough study of Romanian numbers pronunciations, we came to the conclusion that informal pronunciations occur mostly in two digit numbers. Consequently, only two digit numbers will be dealt with in the following subsections.

2.1. Romanian two-digit numbers

Similarly to English, Romanian two-digit numbers are written as single compound words or sequences of simple and compound words (Table 1).

The numbers between 11 and 19 are compound words formed by concatenating the unit words: 1, 2, ..., 9 ("unu", "doi", ..., "nouă" in Romanian), with the preposition "to" ("spre" in Romanian) and with the word "ten" ("zece" in Romanian). Some of these numbers are listed in Table 1. 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 21 and 99 are written as sequences of words obtained 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. Some of these numbers are listed in Table 1. There are no exceptions to this composition rule.

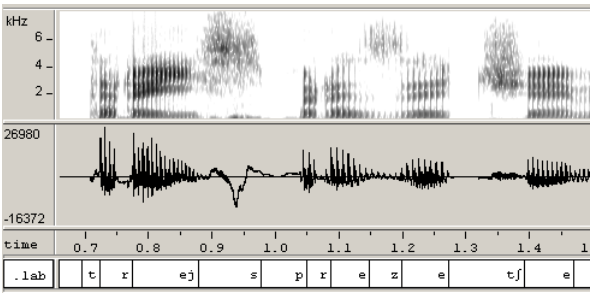
2.2. Pronunciation variations of Romanian numbers and typical ASR errors

In colloquial speech, pronunciations of the Romanian numbers 11 - 19 suffer from the elision phonological phenomena. They are pronounced by changing several syllables (within the compound word) into a shorter one. For example, the word "treisprezece" (13), formally pronounced /trej.spre.ze.ʧe/, is pronounced colloquially /trej.ʃpe/: the syllables /spre/, /ze/ and /ʧe/ have been merged and changed into /ʃpe/. Figure 1 illustrates this behaviour, showing both the formal pronunciation and the informal pronunciation of the number 13.

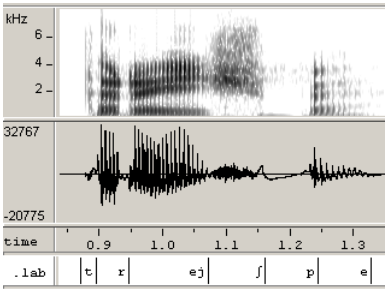
Besides elision, the numbers 17 and 18 suffer from the epenthesis phonological phenomena. For example, the word "optsprezece" (18), formally pronounced /opt.spre.ze.ʧe/, can pronounced colloquially /op.ti.spre.ze.ʧe/, /op.ti.spre.ze.ʧe/, /op.ti.ʃpe/, /op.ti.ʃpe/.

If this phonological phenomenon is not taken into account, a large vocabulary ASR system would erroneously substitute these colloquially-pronounced numbers with acoustic-similar words. Among the most frequent transcription errors made such an ASR system are the following:

- "doisprezece" (12), informally pronounced /doj.ʃpe/, is replaced with
  - "doi şi trei" /doj ʃi trej/ (two and three) or
  - "deci" /deʧʲ/ (consequently) or
  - "doi şi pe" /doj ʃi pe/ (two and on) or
  - "despre ce" /de.spre ʧe/ (about what);



(a) Canonical pronunciation of the number 13



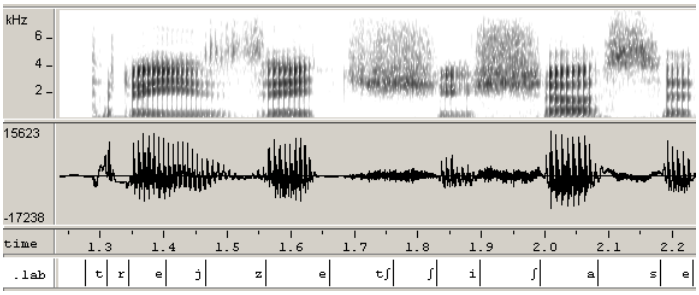
(b) Colloquial pronunciation of the number 13

Figure 1: Pronunciations of 13 (”treisprezece” in Romanian).

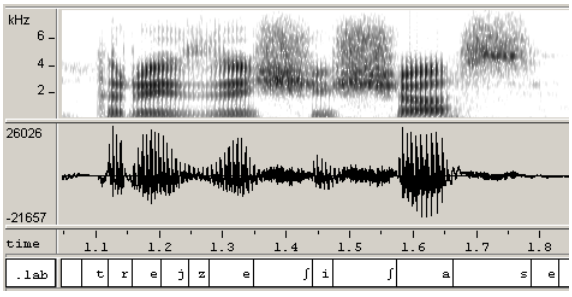
- ”paisprezece” (14), informally pronounced /paj.ʃpe/, is replaced with
  - ”paşte” /paʃ.te/ (Easter) or
  - ”pai şi” /paj ʃi/ (straw and) followed by a word starting with /de/, /te/ or /pe/,
  - ”păi şi” /pəj ʃi/ (well and) followed by a word starting with /de/ or /te/;
- ”cincisprezece” (15), informally pronounced /ʧin.ʃpe/, is replaced with
  - ”cinci pe” /ʧinʧʲ pe/ (five on) or
  - ”cinci că” /ʧinʧʲ kə/ (five that) or
  - ”cinci” /ʧinʧʲ/ (five) followed by a word starting with /pe/ or /de/;
- ”şaisprezece” (16), informally pronounced /ʃaj.ʃpe/, is replaced with ”şi aici” /ʃi aiʧʲ/ (and here) followed by ”de” /de/ (of), ”pe” /pe/ (on) or ”fel” /fel/ (mode);

This type of pronunciation variation can be modelled in an ASR system by adding an additional (colloquial) pronunciation for the words denoting numbers between 11 and 19 in the phonetic dictionary. The solution is quite simple because the elision does not span across several words (it is a within-word pronunciation variant).

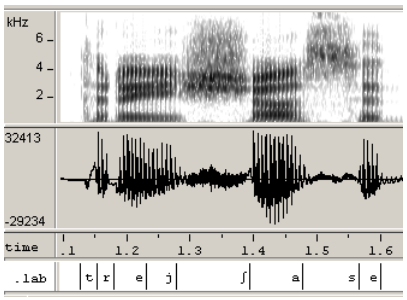
Numbers between 21 - 29, 31 - 39, ..., 91 - 99 are written as three consecutive words (e.g. the number 36 is written ”treizeci şi şase”). In colloquial speech they are pronounced by contracting the three words and eliding one or several syllables. There are two commonly used colloquial pronunciations for these numbers. As an example, the number ”treizeci şi şase” (36), formally pronounced /trej.zeʧ ʲ ʃi ʃa.se/, is pronounced colloquially /trej.ze ʃi ʃa.se/ (the /ʧ/ in the second syllable was elided) or /trej.ʃa.se/ (the second syllable and the second word were elided and the three words contracted). Figure 2 illustrates this behaviour, showing both the formal pronunciation and the informal pronunciations of the number 36.



(a) Canonical pronunciation of the number 36



(b) Colloquial pronunciation #1 of the number 36



(b) Colloquial pronunciation #2 of the number 36

Figure 2: Pronunciations of 36 (”treizeci şi şase” in Romanian).

If this phonological phenomenon is not taken into account, a large vocabulary ASR system would erroneously substitute these colloquially-pronounced numbers with acoustic-similar words. Among the most frequent transcription errors made such an ASR system are the following:

- sequences such as ”douăzeci şi şase” (26), informally pronounced /do.wəʃa.se/ are replaced with ”două şase” /do.wəʃa.se/ (two six),
- sequences such as ”şaptezeci şi trei” (73), informally pronounced /ʃap.teʃ.trej/ are replaced with ”şapte şi trei” /ʃap.te ʃi trej/ (seven and three),
- same errors as above for the sequences starting with ”douăzeci” (20), ”treizeci” (30), etc. followed by any other digit.

Note that the substitutions are very similar to the uttered words from an acoustical point of view. Practically, the acoustic model has done his job: found the correct sequence of acoustic units. It is the phonetic and language models which have not.

As exemplified, these latter phonological phenomena spread across several words. Consequently, the pronunciation variations cannot be added in the phonetic dictionary, because this model stores pronunciations for single words.

Table 2: Examples of corpus modifications made before language modelling to create artificial compound-words for variable pronounceable sequences of words representing numbers

	Phrase in text corpus (Romanian)	**Phrase in text corpus (English)**
Original phrase 1	Evenimentul a avut loc pe douăzeci şi şapte februarie o mie nouă sute nouăzeci şi patru.	The event took place on February twenty-seven nineteen ninety-four.
Modified phrase 1	Evenimentul a avut loc pe *douăzeci_şi_şapte* februarie o mie nouă sute *nouăzeci_şi_patru*.	The event took place on February *twenty-seven* nineteen *ninety-four*.
Original phrase 2	Indicele BET s-a apreciat cu nouă virgulă optzeci şi nouă la sută.	The BET index rose with nine point eighty-nine.
Modified phrase 2	Indicele BET s-a apreciat cu nouă virgulă *optzeci_şi_nouă* la sută.	The BET index rose with nine point *eighty-nine* percent.

2.3. Solution to cross-word pronunciation variation

The modelling solution proposed for cross-word pronunciation variants is not new, but, to the best of our knowledge, it is used for the first time for the Romanian language. The idea is to design artificial compound-words for the sequences of words which have multiple pronunciations and to provide special pronunciations for them in the phonetic dictionary. We propose using the following concrete steps to update the phonetic and the language model of the ASR system:

1. the word sequences for which cross-word pronunciation variation occurs are identified,
2. the word sequences for which the variation was observed are merged into an artificial compound-word,
3. the language model is updated to model these artificial compound-words instead of original sequences of words,
4. the phonetic model is updated to include the various pronunciations for these artificial compound-words.

The above steps are general and could be used for modelling any cross-word pronunciation variations. In this study we focus strictly on colloquially pronounced Romanian numbers and, consequently, the word sequences selected in step #1 are the sequences of three words denoting Romanian numbers 21 - 29, 31 - 39, ..., 91 - 99.

Step #2 involves generating artificial compound words for the selected word sequences by merging the words with an underscore (e.g. "treizeci şi şase" ==> "treizeci_şi_şase").

Step #3 depends on whether the speech recognition system is a rule-grammar ASR system or a large-vocabulary ASR system. In the case of rule-grammar ASR systems, this step involves replacing grammar rules comprising selected word sequences with rules comprising the corresponding artificial compound-words [17]. In the case of large-vocabulary ASR systems, which usually use an n-gram language model (built based on a text corpus), this replacement cannot be done in the model itself. The problem needs to be addressed before language modelling by performing the replacement in the text corpus: the selected sequences of words are replaced with artificial compound-words in the corpus itself. Afterwards, the language model is retrained on the updated text corpus. Eventually, the updated language model will comprise regular words and artificial compound-words.

After in [17] we tackled the informal pronunciation problem with a rule-grammar ASR system, in this study we approached it using a large-vocabulary speech recognition system. Therefore, step #3 involved replacing the selected sequences of words (denoting numbers 21 - 29, 31 - 39, ..., 91 - 99) with artificial compound-words in the training text corpus, followed by retraining the language model. Table 2 exemplifies some text corpus modifications performed.

Note that the proposed method generates an ASR system that outputs regular words and artificial compound-words. In an output post-processing stage, these artificial compound-words have to be transformed into regular words by replacing all underscore characters with spaces.

3. Experimental setup and results

3.1. 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 [18]. The ASR system is built upon the CMU Sphinx speech recognition toolkit [19]. More specifically, the decoding system uses a modified version of the CMU Sphinx 4 Java decoder.

The acoustic models are speaker-independent, 3-state HMMs with output probabilities modelled with GMMs. As speech features we use the classic Mel Frequency Cepstral Coefficients (MFCCs) plus their first and second temporal derivatives (13 MFCCs + deltas + double deltas). The 36 phonemes of the Romanian language are modelled contextually (context dependent phonemes) with 4000 HMM senones. 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 64 GMMs per senone). The acoustic models were trained and optimized on a Romanian read and spontaneous speech corpus.

The baseline language model is a tri-gram, closed-vocabulary (64k words) model, developed with the SRI-LM Toolkit [20] using a text corpus of about 350M words comprising Romanian news articles and TV show transcriptions. The baseline phonetic dictionary contains canonical pronunciations for all the words in the language model.

3.2. Evaluation speech corpus

For evaluating the proposed method we used the same speech corpus as in [17]. This corpus 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 predefined 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. Fourteen speakers recorded partially or entirely 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

Table 3: Evaluation speech corpus

Pronunciation	Utterances	Words	Hours	Speakers
Formal	1150	4949	1.0	12
Informal	1850	7723	1.6	13
Total	3000	12672	2.6	14

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 3.

3.3. Experimental results

Several large-vocabulary ASR setups were evaluated on the speech corpus described above. In all cases, the same acoustic model and speech decoder (described in Section 3.1) were used. The pronunciation and language models differ from one experimental setup to another as follows. The baseline ASR system uses (i) a regular tri-gram language model created using a news text corpus and (ii) a pronunciation model comprising formal pronunciations for all the words in the vocabulary.

The first enhancement of the baseline system involved one improvement: the phonetic dictionary was supplemented with informal pronunciations for the numbers 11 - 19 (only within-word pronunciation variants).

The second enhancement involved two more improvements: (i) the language model was re-trained on the text corpus with word sequences representing numbers replaced by artificial compound-words (as described in Section 2.3 and Table 2) and (ii) the phonetic dictionary was supplemented with colloquial pronunciations for all these artificial compound-words.

The experimental results on the Numbers speech corpus and its two parts ("inf" - informal pronunciations and "for" - formal pronunciations) are presented in Table 4. In the first setup colloquial pronunciations were not modelled at all. In the second setup within-word colloquial pronunciation variations for the two-digit numbers between 11 and 19 were inserted in the pronunciation model (these numbers are written with single words, see Table 1). Finally, in the third setup, the language model was updated to model two-digit numbers between 21 and 99 as artificial compound-words (these numbers are normally written with several words, see Table 1) and cross-word informal pronunciation variations (for these compound-words) were inserted into the pronunciation model.

As Table 4 shows, the pronunciation modelling approaches proposed in this paper have significant beneficial effects for the task of recognizing informally pronounced numbers in spontaneous speech. Within-word pronunciation modelling of numbers brings a relative WER improvement of 12% over the baseline, while both within-word and cross-word pronunciation modelling of numbers bring a relative WER improvement of 35% over the baseline. 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.

4. Discussion and error analysis

Following the experiment, a detailed error analysis was performed to better understand what worked and what did not work in every ASR setup. The target was to answer questions such as

- Which systematic errors were solved and which were not?
- Why modelling informal pronunciations also helped improve the accuracy on formally pronounced numbers?

To answer these questions, the ASR errors for the words representing the numbers 11, 12, ... 19 and 20, 30, ..., 90 were listed, counted and analysed.

Most of the errors for the informally-pronounced numbers between 11 and 19 were successfully solved. Overall, out of 172 occurrences of such words, there were 137 errors in the baseline and only 29 errors after adding informal pronunciations for these numbers in the phonetic dictionary. There is, though, one exception: for "şaisprezece" (16), informally pronounced /ʃaj.ʃpe/, only half of the 32 errors made by the baseline system were solved. Most of the substitutions with "şi aici" /ʃi aitʃʲ/ (and here) followed by "de" /de/ (of), "pe" /pe/ (on) or "fel" /fel/ (mode) were not solved.

Another interesting observation is that one speaker systematically utters "doisprezece" (12) using an informal pronunciation which was not envisioned: /doj.spre.e.ʧe/ (the /z/ in the third syllable is elided). This triggered all ASR systems to replace "doisprezece" with "despre ce" /de.spre ʧe/ (about what).

The introduction of alternative, informal pronunciations for the numbers 11-19 triggered new ASR errors. The introduction of /ʧin.ʃpe/, /ʃaj.ʃpe/ and /no.wɔ.ʃpe/, as alternative pronunciations for "cincisprezece" (15), "şaisprezece" (16) and "nouăsprezece" (19) introduced 46 new errors. These numbers were output by the ASR system instead of other 2-digit formally-pronounced numbers such as "cincizeci şi ..." (fifty-...), "şaizeci şi ..." (sixty-...), "nouăzeci şi ..." (ninety-...). The bright part was that these errors were eventually resolved by the introduction of pronunciation variants for these latter sequences of words (in the second system enhancement).

In conclusion, modelling informal pronunciations for the numbers 11-19 solved most of the ASR errors for these numbers, but introduced new errors, which were eventually all solved by modelling informal pronunciations for the numbers 21-29, 31-39, 91-99 also.

Most of the errors for the informally-pronounced numbers between 21-29, 31-39, 91-99 were also successfully solved. Overall, out of 886 occurrences of such sequences of words, there were 429 errors in the baseline and 87 errors after adding artificial compound words in the language models and their informal pronunciations in the phonetic dictionary. 75% of the remaining errors are for 2-digit numbers starting with 5, 6 and 8 (fifty-..., sixty-..., and eighty-...). No systematic problem was identified by analysing these remaining errors.

Table 4: Experimental results obtained for transcribing the two parts (comprising informal and formal pronunciations) of the Numbers speech corpus

Language Model	**Pronunciation Model**	**WER[%]**		
		inf	**for**	**all**
Regular words	Formal pronunciations only	32.5	11.3	24.2
Regular words	+ within-word informal pronunciations	28.2	10.5	21.3
+ 21-99 modelled as compound words	+ cross-word informal pronunciations	19.7	9.2	15.6

Besides the above errors, the final ASR system also makes an important number of systematical errors such as:

1. Insertion of "lei" /lej/ (Romanian national currency) after a sequence of words representing a number or substitution of "mii" /mij/ (thousands) with "lei". This is due to the fact that the probabilities of such word sequences (as modelled by the LM) are high.
2. Substitution of "opt sute" /opt sute/ (eight hundreds) with "o sută" /o sutə/ (one hundred). This is due to the high acoustic similarity of the two sequences and the fact that the phones /p/ and /t/ are very short and voiceless.
3. Substitution of "mii" /mij/ (thousands) with "miei" /miej/ (lambs), for the same reason as above.

Modelling informally-pronounced numbers between 21-29, 31-39, 91-99 also triggered an improvement in WER on the formally-pronounced part of the Numbers speech corpus. By analysing the avoided errors, we concluded that 2-digit numbers starting with 5 (fifty-...) were (wrongly) pronounced informally in 67% of the cases in this part of the corpus also. This explains the small drop in WER for "formally-pronounced" numbers.

Another interesting observation was made by analysing the errors made by the system on formally-pronounced numbers. In 20% of the 69 cases, when trying to pronounce formally the word "cincizeci" /ʧinʧʲ.zeʧʲ/ (fifty), the speakers made a longer pause between syllables, probably due to the difficulty of pronouncing /ʧʲ/ followed by /z/. This triggered the ASR system to output the sequence "cinci zeci" (five tens) or "cinci zece" (five ten) instead of "cincizeci" (fifty).

5. Conclusion

This work presented the problem of automatically transcribing speech comprising colloquial pronunciations of Romanian numbers. A technique for modelling these pronunciations in a large-vocabulary ASR system was proposed. The pronunciation modelling technique was evaluated on a speech corpus comprising rational numbers pronounced formally and informally by 14 speakers. The experiments showed a relative WER improvement of 12% over the baseline when within-word pronunciation variations are taken into account and a relative WER improvement of 35% when cross-word pronunciation variations are also modelled. The errors made by the baseline ASR system and by the enhanced system were analysed, compared and discussed.

6. Acknowledgement

This work was supported in part by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, throught project no. 25/2014 and in part by the Sectoral Operational Programme "Human Resources Development" 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU /159/1.5/S/134398.

7. References

[1] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary continuous speech recognition," in *Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, USA*, vol. 1, 1996, pp. 6–9.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[3] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, vol. 29, no. 2, pp. 83–98, 1999.

[4] J. M. Kessens, M. Wester, and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 193–207, 1999.

[5] D. AbuZeina, W. Al-Khatib, M. Elshafei, and H. Al-Muhtaseb, "Cross-word Arabic pronunciation variation modeling for speech recognition," *International Journal of Speech Technology*, vol. 14, no. 3, pp. 227–236, 2011.

[6] G. Perennou and L. Pousse, "Dealing with pronunciation variants at the language model level for automatic continuous speech recognition of French," in *Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece*, 1997, pp. 2727–2730.

[7] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks," *Speech Communication*, vol. 27, no. 1, pp. 63 – 73, 1999.

[8] M.-Y. Tsai, F.-C. Chou, and L. shan Lee, "Pronunciation modeling with reduced confusion for mandarin chinese using a three-stage framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 661–675, Feb 2007.

[9] P. Karanasou and L. Lamel, "Comparing SMT methods for automatic generation of pronunciation variants," in *Advances in Natural Language Processing*. Springer Berlin Heidelberg, 2010, pp. 167–178.

[10] H. van den Heuvel, B. Réveil, and J. Martens, "Pronunciation-based ASR for names," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 2991–2994.

[11] M. Wester, "Pronunciation modeling for ASR - knowledge-based and data-derived methods," *Computer Speech & Language*, vol. 17, no. 1, pp. 69–85, Jan 2003.

[12] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 327–332, May 2001.

[13] M. Adda-Decker, P. Boula de Mareüil, and L. Lamel, "Pronunciation variants in french: schwa & liaison," *International Congress of Phonetic Sciences*, pp. 2239–2242, 1999.

[14] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition." in *Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece*, 1997, pp. 2379–2382.

[15] E. Giachin, A. Rosenberg, and C.-H. Lee, "Word juncture modeling using phonological rules for HMM-based continuous speech recognition," *Computer Speech & Language*, vol. 5, no. 2, pp. 155–168, Apr 1991.

[16] K. Ries, F. D. Buo, and A. Waibel, "Class phrase models for language modeling," in *Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, USA*, vol. 1, 1996, pp. 398–401.

[17] H. Cucu, A. Caranica, A. Buzo, and C. Burileanu, "On transcribing informally-pronounced numbers in romanian speech," in *Proceedings of 38th International Conference on Telecommunications and Signal Processing, Prague, Czech Republic*, 2015, accepted for publication.

[18] H. Cucu, A. Buzo, L. Petrică, D. Burileanu, and C. Burileanu, "Recent improvements of the SpeeD Romanian LVCSR system," in *Proceedings of the 10th International Conference on Communications, Bucharest, Romania*, 2014, pp. 1–4.

[19] P. Lamere, P. Kwok, W. Walker, E. Gouvêa, R. Singh, B. Raj, and P. Wolf, "Design of the CMU Sphinx-4 decoder," in *Proceedings of the 8th European Conference on Speech Communication and Technology, Geneve, Switzerland*, 2003, pp. 1181–1184.

[20] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.