Dealing with Numbers in Grapheme-Based Speech Recognition*

Miloš Janda, Martin Karafiát, and Jan Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence
Czech Republic
{ijanda,karafiat,cernocky}@fit.vutbr.cz

Abstract. This article presents the results of grapheme-based speech recognition for eight languages. The need for this approach arises in situation of low resource languages, where obtaining a pronunciation dictionary is time- and cost-consuming or impossible. In such scenarios, usage of grapheme dictionaries is the most simplest and straight-forward. The paper describes the process of automatic generation of pronunciation dictionaries with emphasis on the expansion of numbers. Experiments on GlobalPhone database show that grapheme-based systems have results comparable to the phoneme-based ones, especially for phonetic languages.

Keywords: LVCSR, ASR, grapheme, phoneme, speech recognition.

1 Introduction

With fast spread of speech processing technologies over the last decade, there is a pressure to speech processing community to build Large Vocabulary Continuous Speech Recognition (LVCSR) systems for more and more different languages. One of essential components in the process of building speech recognizer is pronunciation dictionary, that maps orthographic representation into a sequence of phonemes — the sub words units, which we use to define acoustic models during the process of training and recognition.

The acquisition of quality hand-crafted dictionary requires linguistic knowledge about target languages and is time- and money-consuming, especially for rare and low-resource languages. For these, several approaches for automatic or semi-automatic generation of dictionaries have been introduced, typically based on contextual pronunciation rules [1], neural networks [2] or statistical approaches [3].

The most straightforward method is to generate pronunciation dictionary as sequence of graphemes and thus to directly use orthographic units as acoustic models (see [4,5]). This approach is suitable for phonetic languages, where relation between the written and the spoken form is reasonably close. The most widely used phonographic writing script is the Roman script, so it is not surprising, that grapheme-based speech recognition (GBSR) has been extensively tested on Western languages using this script. Later

* This work was partly supported by Czech Ministry of Trade and Commerce project No. FR-TI1/034, by Czech Ministry of Education project No. MSM0021630528 and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 438–445, 2012.

Table 1. Numbers of speakers, amounts of audio material (hours) and sizes of dictionary (words)

Lang.	Speakers	TRAIN (h)	TEST (h)	DICT
CZ	102	27	1.9	33k
EN	311	15	1.0	10k
GE	77	17	1.3	47k
PO	102	27	1.0	56k
SP	100	21	1.2	42k
RU	115	20	1.4	29k
TU	100	15	1.4	33k
VN	129	16	1.3	8k

experiments and results in this paper show, that the grapheme-based approach is also suitable for Cyrillic [6] or for the tonal languages like Vietnamese or Thai [7].

2 Experimental Setup

This section presents the data corpus and details the generation of grapheme based dictionaries with two possibilities (with and without expansion of numbers).

2.1 Data

GlobalPhone [8] was used in our experiments. The database covers 19 languages with an average of 20 hours of speech from about 100 native speakers per language. It contains newspaper articles (from years 1995–2009) read by native speakers (both genders). Speech was recorded in office-like environment by high quality equipment. We converted the recordings to 8 kHz, 16 bit, mono format.

The following languages were selected for the experiments: Czech (CZ), German (GE), Portuguese (PO), Spanish (SP), Russian (RU), Turkish (TU) and Vietnamese (VN). These languages were complemented with English (EN) taken from Wall Street Journal database. See Table 1 for detailed numbers of speakers, data partitioning and vocabulary sizes. Each individual speaker appears only in one set. The partitioning followed the GlobalPhone recommendation (where available).

When preparing the databases for baseline phoneme-based systems, several problems were encountered. The biggest issue was the low quality of dictionaries with many missing words. The Vietnamese dictionary was missing completely. The typos and miss-spelled words were corrected, numbers and abbreviations were labeled and missing pronunciations were generated with an in-house grapheme-to-phoneme (G2P) tool trained on existing pronunciations from given language. The dictionaries for Vietnamese and Russian were obtained from Lingea[1]. The CMU dictionary[2] was used for English. Each language has its own phoneme set and for better handling with different locales, all transcripts, dictionaries and language models (LMs) were converted to Unicode (UTF-8).

[1] `http://www.lingea.com`

[2] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Table 2. OOV rates, dictionary sizes, LM sizes and sources for individual languages

Lang	OOV	LM Dict Size	LM Corpus Size	WWW Server
CZ	3.08	323k	7M	www.novinky.cz
EN	2.30	20k	39M	WSJ - LDC2000T43
GE	1.92	375k	19M	www.faz.net
PO	0.92	205k	23M	www.linguateca.pt/cetenfolha
SP	3.10	135k	18M	www.aldia.cr
RU	1.44	485k	19M	www.pravda.ru
TU	2.60	579k	15M	www.zaman.com.tr
VN	0.02	16k	6M	www.tintuconline.vn

The data for LM training were obtained from Internet newspaper articles using `RLAT` and `SPICE` tools from the KIT/CMU[3]. The sizes of corpora gathered for LM training, and the sources are given in Tab. 2. Bigram LMs were generated for all languages except Vietnamese — a syllable language — for which a trigram LM was created.

2.2 Grapheme-Based Dictionaries

As proposed in the Introduction, the conversion of dictionaries to grapheme form was done. Word lists were obtained from current pronunciation dictionaries. An alternative would be to derive lists of words directly from transcripts, but we wanted to guarantee the same size of vocabulary in both (phoneme and grapheme) dictionaries and thus guarantee the same OOV rate for both systems and comparable results.

Prior to dictionary conversion to grapheme form, the word-lists were pre-processed: special characters like asterisk, brackets, colons, dashes, dollar symbols, etc. were removed. In the first version of grapheme dictionaries, we also removed all marked numbers from the vocabulary. After these operations, the grapheme based dictionary was obtained by simple splitting the words to letters, and finally, all graphemes were converted to lowercase (e.g. WORD → w o r d).

The transcripts of CZ, EN, VN did not contain any numbers, but we had to investigate how to deal with them for GE, SP, PO, RU, and TU. With deletion of numbers from dictionaries, we had to adequately change the transcripts to be consistent. One option was to remove all utterances, where a number is spoken (*grap_v0*). Another option was to map missing numbers in transcript into "unknown" <*UNK*> symbol (*grap_v1*).

The above mentioned processing of numbers however led to significant loss of acoustic data available for training (see Table 3). In average, we lose about 3.4 hours of data for the first variant, which represents about 17% on 20 hours of speech. The rate of numbers in the original dictionaries is about 3%. These differences can produce large degradation of recognition accuracy in the final results, so we decided to make another versions of grapheme dictionaries using numbers expansion.

[3] `http://i19pc5.ira.uka.de/rlat-dev`, `http://plan.is.cs.cmu.edu/Spice`

Table 3. Amount of audio data in different setups (with and without numbers)

	With numbers		Without numbers		Difference
Lang	[hours]	[utts]	[hours]	[utts]	[hours in %]
GE	16.37	9034	14.96	8390	-8.6%
PO	16.75	7350	12.33	5805	-26.3%
SP	15.36	5227	10.77	4064	-29.8%
RU	19.49	9771	16.73	8822	-14.1%
TU	14.49	5988	10.75	4775	-25.8%

2.3 Grapheme-Based System with Number Expansion

From the previous analysis, it is obvious that numbers need to be processed in a less aggressive way. Then second version of dictionaries (*grap_v2*) with number expansion were generated. For number expansion we used standard ICU library[4], which can be used for most languages and supports large variety of locales. With number expansion, we obtained complete dictionaries with all words including numbers and all acoustic data, without any loss of information, could be used.

We observed that a number in dictionary can have two meanings. One as normal word — cardinal number (e.g. 911 → n i n e h u n d r e d a n d e l e v e n), and another as a sequence of digits, i.e. for phone numbers, credit card numbers, etc. (e.g. 911 → n i n e o n e o n e). In *grap_v2* version, we did not use any variants and transcribed numbers in the first mentioned way (as cardinal number, e.g. 911 → n i n e h u n d r e d a n d e l e v e n).

Then, another version of dictionaries was produced (*grap_v3*), where we combined both of these variants and all numbers were expanded as in version *grap_v2* plus as a sequence of digits (so each number exists in dictionary two times with different pronunciations). In fact, pronunciation of numbers in form of single spoken digits is not frequent in GlobalPhone data, so we did not expect any substantial improvement with *grap_v3*.

2.4 Number Expansion for Spanish with Gender Dependency

For Spanish, a more sophisticated expansion of numbers was tested. Here, the gender of noun following a number can change the expansion of the number, for example:

```
un lápiz (one pencil)
una pluma (one pen)
uno (one - as single number)

cincuenta y un lápices (fifty-one pencils)
cincuenta y una plumas (fifty-one pens)
cincuenta y uno (fifty-one - as single number)
```

[4] http://site.icu-project.org/

doscientos dos coches (202 cars)
doscientas dos casas (202 houses)

The underlined numbers vary according to gender. When a number ends in '-uno' (one), the form '-un' is used before masculine nouns, and '-una' before feminine nouns. The 'uno' form is used only in counting. The hundreds portions of numbers change in gender even when other parts of the number intervene before the noun[5].

The limitation of this expansion is in fact, that whole transcription is needed for dictionary generation, as we first need to obtain pairs of numbers and corresponding nouns. With such a list of pairs, we can expand numbers in correct way, according to noun gender. As will be seen in the results, this method significantly improves accuracy, even over results of phoneme-based system. On the other hand, this approach uses morphology information about target language, which can be also used in standard phoneme dictionary and thus accuracy of phoneme-based system could also be improved.

3 Experimental Framework

The KALDI toolkit[6] was used for all recognition experiments [9]. We setup five systems:

- **Phon**: phoneme-based, which is set as a baseline.
- **Grap_v0** - grapheme-based, without numbers (with reduced acoustic data)
- **Grap_v1** - grapheme-based, without numbers (no reduction of data, numbers mapped to <*UNK*> symbol in transcripts)
- **Grap_v2** - grapheme-based with expanded numbers (no reduction of data). All numbers expanded as cardinal number
- **Grap_v3** - grapheme-based with expanded numbers (no reduction of data). All numbers expanded in two meanings — as cardinal number and in form of single spoken digits.

As features, we extract 13 Mel-frequency cepstral coefficients (MFCCs) and compute delta and delta-delta features. For all four setups, we first train a monophone system (*mono*) with about 10k diagonal Gaussians. Next, we train initial triphone system with about 50k diagonal covariance Gaussians (5000 states). This system is retrained into triphone system (*tri2c*) with the same number of parameters, and per-speaker cepstral mean normalization applied. The last system — *SGMM* — is built on subspace-GMMs [10] modeling of triphones. This system uses 400-component full-covariance Gaussian background model, about 6,000 tree leaves and 22k Gaussians in total.

4 Results

All results are given in terms of word accuracy. Table 4 presents the results for monophone system, the second column shows numbers of phonemes, resp. graphemes for different languages. Last column gives absolute improvement in accuracy between phoneme (phon) and grapheme system (grap_v2).

[5] http://spanish.about.com/cs/forbeginners/a/cardinalnum_beg.htm
[6] http://kaldi.sourceforge.net

Table 4. Accuracy of monophone system for different languages

	Count	MONO					ACC (Diff)
Lang	phon/grap	phon	grap_v0	grap_v1	grap_v2	grap_v3	phon/grap_v2
CZ	41/44	**64.2**	62.7				−1.5%
EN	40/27	**71.1**	43.9				−27.2%
GE	42/31	**51.9**	42.8	42.2	43.1	43.3	−8.8%
PO	34/40	**54.1**	48.0	47.6	48.3	47.7	−5.8%
SP	36/34	**61.5**	58.5	59.7	59.5	59.6	−2.0%
RU	54/34	**50.5**	47.1	47.4	47.3	47.2	−3.2%
TU	30/33	46.9	46.4	48.0	47.1	**48.1**	0.2%
VN	85/94	**61.1**	55.7				−4.2%

Table 5. Accuracy of triphone GMM system for different languages

	TRI2c					ACC (Diff)
Lang	phon	grap_v0	grap_v1	grap_v2	grap_v3	phon/grap_v2
CZ	**76.0**	75.9				−0.1%
EN	**82.6**	76.0				−6.6%
GE	**71.0**	70.2	70.5	70.7	70.8	−0.3%
PO	**72.9**	70.3	69.5	71.8	71.8	−1.1%
SP	75.4	74.5	**75.6**	75.4	75.4	0%
RU	**65.2**	63.3	63.9	64.1	64.1	−1.1%
TU	66.0	63.9	65.7	**66.1**	**66.1**	0.1%
VN	71.1	**71.6**				0.5%

As we can see, baseline phoneme-based systems have the best results in monophone training for almost all languages, the grapheme-based systems are about 2–8% absolutely worse. Only Turkish is an exception, with 0.2% better accuracy. On the other hand, the biggest hit is observed for English. Here, the results and big reduction of the number of acoustic units (from 40 phonemes to 27 graphemes) are related to the fact, that English spelling is not phonetically-based.

Table 5 shows the results for triphone GMM system. Here, grapheme-based setups have about 0.1–2% worse accuracy than phonemes, for EN, the degradation is about 6% against the baseline. These improvements are caused by possibility of triphone system to model wider context of graphemes. For some languages (SP, TU, VN), triphone grapheme based system works even better than phoneme one, this fact could indicate poor quality of the original dictionaries.

Table 6 presents the results for SGMM systems. Although we did not trained jointly shared parameters on all languages, we could see improvement in accuracy, obtained by simple usage of Sub-space Gaussian Mixture Models on each language. SGMMs have about 3–6% better results for phonemes and about 4–7% for graphemes against similar systems in triphone GMM training. The gap between phoneme and grapheme systems is further decreased.

The last Table 7 summarizes the result for Spanish, where three columns represent different systems (monophone, triphone GMM, SGMM) and rows present the results

Table 6. Accuracy of SGMM system for different languages

	SGMM					ACC (Diff)
Lang	phon	grap_v0	grap_v1	grap_v2	grap_v3	phon/grap_v2
CZ	**79.0**	78.5				−0.5%
EN	**85.4**	79.7				−5.7%
GE	**76.2**	75.5	75.1	75.6	75.1	−0.6%
PO	**76.3**	74.2	74.8	75.6	75.8	−0.7%
SP	78.7	78.6	78.9	**79.4**	79.2	0.7%
RU	68.5	68.2	68.4	**69.1**	68.3	0.6%
TU	70.3	69.7	70.3	70.0	**70.8**	−0.3%
VN	77.9	**78.6**				0.7%

Table 7. Accuracy of SGMM system for different languages

SP	MONO	TRI2c	SGMM
Phon	**61.5**	75.4	78.7
Grap_v2	59.5	75.4	79.4
Grap_v4	59.7	**76.0**	**79.6**

of baseline phoneme system, grapheme system with number expansion (grap_v2) and results after gender dependent expansion of numbers (grap_v4) as described in section 2.4.

Advanced expansion outperforms baseline in triphone GMM (0.6% absolute better) and SGMM system (+0.2%). Monophone result is still under baseline, but there is an improvement over the system with basic expansion of numbers (grap_v1), thus we can claim, that improved expansion of number gives better results than basic one. On the other hand, for rare languages we mostly know nothing about language morphology and rules, so advanced expansion of numbers is out of question.

5 Conclusion

We have shown that grapheme-based speech recognition, that copes with the problem of low-quality or missing pronunciation dictionaries, is applicable for phonetic languages and also tonal languages like Vietnamese. For class of non-phonetic languages, like English, using of models with wider context gives also comparable results and grapheme based approach can be, with small limitation, usable also for these languages. Improved expansion of numbers for Spanish also answered the question, whether we are able to do the number expansion better in situation, where we have information about target language and its morphology and rules. Grapheme-based straightforward approach, supported by the expansion of numbers in dictionaries, is advantageous especially in situation of low-resource languages and could be successfully used in building speech recognizers for rare languages.

References

1. Black, A., Lenzo, K., Pagel, V.: Issues in building general letter to sound rules. In: Proceedings of the ESCA Workshop on Speech Synthesis, Australia, pp. 77–80 (1998)
2. Fukada, T., Sagisaka, Y.: Automatic generation of multiple pronunciations based on neural networks. Speech Communication 27(1), 63–73 (1999)
3. Besling, S.: Heuristical and statistical Methods for Grapheme-to-Phoneme Conversion, Konvens, Wien, Austria, pp. 23–31 (1994)
4. Killer, M., Stüker, S., Schultz, T.: Grapheme Based Speech Recognition. In: Proceedings of the EUROSPEECH, Geneve, Switzerland, pp. 3141–3144 (2003)
5. Schillo, C., Fink, G.A., Kummert, F.: Grapheme Based Speech Recognition For Large Vocabularies. In: Proceedings of ICSLP 2000, pp. 129–132 (2000)
6. Stüker, S., Schultz, T.: A Grapheme Based Speech Recognition System for Russian. In: Specom 2004 (2004)
7. Charoenpornsawat, P., Hewavitharana, S., Schultz, T.: Thai grapheme-based speech recognition. In: Proceedings of the Human Language Technology Conference of the NAACL, Stroudsburg, PA, USA, pp. 17–20 (2006)
8. Schultz, T., Westphal, M., Waibel, A.: The globalphone project: Multilingual lvcsr with janus-3. In: Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzeň, Czech Republic, pp. 20–27 (1997)
9. Povey, D., Ghoshal, A., et al.: The Kaldi Speech Recognition Toolkit. In: Proceedings of the ASRU, Hawaii, US (2011)
10. Povey, D., Burget, L., et al.: The subspace Gaussian mixture model – A structured model for speech recognition. Computer Speech and Language 25(2) (2011)