GRAPHEME BASED SPEECH RECOGNITION

Miloš Janda
Doctoral Degree Programme (2), FIT BUT
E-mail: xjanda16@stud.fit.vutbr.cz

Supervised by: Martin Karafiát and Jan Černocký
E-mail: {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: speech recognition, LVCSR, ASR, grapheme, phoneme, low-resource languages

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 languages, several approaches for automatic or semi-automatic generation of dictionaries have been introduced. These methods are 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 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 8kHz, 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 Tab. 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).

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

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

When preparing the databases for baseline phoneme-based system, 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. The CMU dictionary 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).

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

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

The data for LM training were obtained from Internet newspaper articles using `RLAT` and `SPICE` tools from the KIT/CMU. 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 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 (*grap_v0*), 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 $\rightarrow$ 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.

	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 %

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

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, 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 (e.g. 911 $\rightarrow$ n i n e h u n d r e d a n d e l e v e n), and second as a sequence of digits, i.e. for phone numbers, credit card numbers, etc. (e.g. 911 $\rightarrow$ n i n e o n e o n e). In fact, this situation affects only tiny percentage of numbers so we did not use any variants and transcribed numbers in the first mentioned way (as cardinal numbers, e.g. 911 $\rightarrow$ n i n e h u n d r e d a n d e l e v e n).

3 EXPERIMENTAL FRAMEWORK

The KALDI toolkit was used for all recognition experiments [9].

We setup four systems:

- **Phon**: phoneme-based, which is set as a baseline.
- **Grap_v0** - grapheme-based, without numbers (with reduced lists of 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).

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.

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.

	Count	MONO				ACC (Diff)
Lang	phon/grap	phon	grap_v0	grap_v1	grap_v2	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	-8.8 %
PO	34/40	**54.1**	48.0	47.6	48.3	-5.8 %
SP	36/34	**61.5**	58.5	59.7	59.5	-2.0 %
RU	54/34	**50.5**	47.1	47.4	47.3	-3.2 %
TU	30/33	46.9	46.4	**48.0**	47.1	0.2 %
VN	85/94	**61.1**	55.7			-4.2 %

Table 4: Accuracy of monophone system for different languages.

As we can see, the baseline phoneme-based system has the best results in monophone training for almost all languages, the grapheme-based systems are about 2 - 8% absolutely worse. The biggest gap is observed for English, where the results and big reduction of the number of acoustic units (from 40 phonemes to 27 graphemes) are related to the fact, that English is not phonetic language.

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 the 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.

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 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 this class of non-phonetic languages. This 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.

	TRI2c				ACC (Diff)
Lang	phon	grap_v0	grap_v1	grap_v2	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	-0.3 %
PO	**72.9**	70.3	69.5	71.8	-1.1 %
SP	75.4	74.5	**75.6**	75.4	0 %
RU	**65.2**	63.3	63.9	64.1	-1.1 %
TU	66.0	63.9	65.7	**66.1**	0.1 %
VN	71.1	**71.6**			0.5 %

Table 5: Accuracy of triphone GMM system for different languages.

ACKNOWLEDGMENTS

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).

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, 1998, pp.77–80

[2] Fukada, T., Sagisaka, Y.: Automatic generation of multiple pronunciations based on neural networks. Speech Communication, Volume 27, Issue 1, 1999, p. 63-73

[3] Besling, S.: Heuristical and statistical Methods for Grapheme-to-Phoneme Conversion, Konvens, Wien, Austria, 1994, p.23-31

[4] Killer, M., Stüker, S., Schultz, T.: Grapheme Based Speech Recognition. In Proceedings of the EUROSPEECH, Geneve, Switzerland, 2003, pp. 3141-3144

[5] Schillo, Ch., Fink, G. A., Kummert, F.: Grapheme Based Speech Recognition For Large Vocabularies. In Proceedings of ICSLP '00, 2000, p. 129-132

[6] Stüker, S., Schultz, T.: A Grapheme Based Speech Recognition System for Russian. 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, 2006. p. 17-20

[8] Schultz, T., Westphal, M., Waibel, A.: The globalphone project: Multilingual lvcsr with janus-3. In in Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, Czech Republic, 1997, p. 20-27

[9] Povey, D., Ghoshal, A., et. al: The Kaldi Speech Recognition Toolkit, In Proceedings of the ASRU, Hawaii, US, 2011