AUTOMATIC GENERATION OF PRONUNCIATION DICTIONARIES BASED ON DIARIZATION

Miloš Janda
Doctoral Degree Programme (3), FIT BUT
E-mail: xjanda16@stud.fit.vutbr.cz

Supervised by: Jan Černocký and Martin Karafiát
E-mail: {cernocky,karafiat}@fit.vutbr.cz

Abstract: This paper presents the results of eight speech recognizers with automatically generated pronunciation dictionaries. Diarization approach, typically used in speaker recognition, could be modified for purpose of automatic generation of dictionaries, targeting the low resource languages where acquisition of hand-crafted pronunciation dictionary is time- and cost-consuming or impossible. Experiments on GlobalPhone database show that diarization approach is suitable alternative for grapheme-based method.

Keywords: speech recognition, LVCSR, vocabulary, pronunciation dictionary, lexicon, diarization

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 are used 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 automatic or semi-automatic generation of dictionaries several approaches have been introduced, typically based on contextual pronunciation rules [2], neural networks [3] or statistical approaches [1]. The most straightforward method for automatic dictionary generation is to model pronunciation dictionary as sequence of graphemes and thus to directly use orthographic units as acoustic models. This grapheme-based method was presented in [5].

Finally, idea of using diarization approach for automatic segmentation to find acoustic sub-words units [4], was used in our experiments and is discussed in section 3.

2 EXPERIMENTAL SETUP

2.1 DATA

GlobalPhone [6] 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 (www.lingea.com). 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). Pre-segmentation was done using phoneme recognizer (recordings were divided into speech and non-speech parts).

Transcripts of training part of data were used for generation of standard LMs. Bigram LMs were obtained for all languages except Vietnamese — a syllable language — for which a trigram LM was created.

3 DIARIZATION APPROACH

Diarization technique is closely related to speaker recognition, but could be easily modified to task of generation sub-words units. Originally diarization is used to divide multi-speaker recordings into segments according to particular speaker. In our experiments, diarization is used to detect speaker-independent sub-word units, where it is utilized to automatically generate pronunciation dictionaries in combination with G2P system.

3.1 Automatic generation of pronunciation dictionaries

Automatic generation of pronunciation dictionaries using diarization method starts from grapheme-based triphone GMM system with Cepstral Mean Normalization applied and consists of two parts:

- **segmentation/"cutting"** part — task is to divide acoustic data into small chunks (units) according the BIC (Bayesian Information Criterion) in diarization technique.
- **clustering/"merging"** part, where task is to determine which chunks (units) are the same (belong to one acoustic unit) and label them.

The finding and clustering segments from one speaker is based on the BIC, which is used to determine the border (edge) in recording, where speaker is changed and thus segment utterance into speaker-dependent parts. On the same base, but with smaller frame length, we could use the BIC to find the best division points in utterance (or in each word), where the acoustic change is highest, and thus define the sub-words units. Then this sub-words units (chunks) are clustered into final set of independent acoustic units.

Generally, let $\mathbf{X} = \{\mathbf{x_i}, i = 1, \ldots N\}$ be the feature vectors of input speech; let M be the candidates of desired parametric models. The BIC criterion is defined as:

$$BIC(M) = \ln L(X,M) - \lambda \frac{\#M}{2} \log N$$

Where $L(X,M)$ is the likelihood of input speech given the model of M, and $\#M$ is the number of parameters in the model M; N is the sample size of input speech.

The process of diarization is recursively running on each word. In the first iteration we have whole recording (one word). All possible divisions are tested and the BIC is computed. At a point, where the BIC is highest, the utterance is cut in two segments. Then second iteration is performed in each part and segments are recursively divided, until process of segmentation is done. Stopping criterion could be one or, as in our case, combination of all states:

- The total number of segments; if reach some defined number (e.g. length of word in graphemes).
- The BIC value is smaller than some threshold; no new division point was found and recursion is stopped in current segment.
- Length of segment is not long enough and could not be again divided.

After segmentation to chunks, the second part of process is performed. All chunks are clustered and labeled, in our examples by the nearest grapheme. Examples of original phoneme dictionary and dictionary based on sub-words units, are given in the next two blocks. Each variant of dictionary is presented for Czech and German language.

CZ_phon

BANKA	b a N k a
BŘEZEN	b rZ E z E n
CENNĚJŠÍ	tS E J E j S i:
VÝKONNÉ	v i: k o n E:

GE_phon

ABGESAGT	A p g _at z a: k t_d
DÜRRE	d_d y _3_backs _at
IMMER	I m _3_backs
ZIEHT	ts i: t_d

CZ_unit

BANKA	b a n k a
BŘEZEN	b ř e z e n
CENNĚJŠÍ	c e n ě j š í
VÝKONNÉ	v ý k o n é

GE_unit

ABGESAGT	a b g e s a g t
DÜRRE	ü r e
IMMER	i m e r
ZIEHT	z i e h

3.2 EXPERIMENTAL FRAMEWORK

Experiments were done on eight languages (listed in table 1). For segmentation we used simple diarization system based on the GMM and the BIC. As features were used 19-dimensional MFCCs vectors. Models had 5 Gaussians with diagonal covariance matrices. For our experiments, we adopted diarization system, originally developed for NIST Speaker Recognition evaluations.

Finally, we setup five LVCSR systems:

- **Phon_v1** - phoneme-based, set as a baseline.
- **Grap_v1** - grapheme-based, without numbers (no reduction of data, numbers mapped to $<UNK>$ symbol in transcripts)

- **Units_v1** - based on automatically generated units, labeled by graphemes, big OOV rate (about 12%)
- **Units_v1x** - added simple automatic cleaning of wrong pronunciation (checking grapheme/units ratio). Missing words were generated via G2P (OOV rate around 1%)
- **Units_v1x_G2P** - trained G2P on *Units_v1x* is used for generation of completly new dictionary. G2P is there in role of “cleaner”, because only one the best pronunciation variant is selected for each word.

4 RESULTS

All results are given in terms of word accuracy. Table 2 presents the results of monophone versions for each system. As we can see, for all languages the accuracy of the first unit system (*units_v1*) is very low, about 20-23% worse than baseline. Very simple cleaning and generation of missing words gave us 5-10% absolute in the second unit system (*units_v1x*) and additional 2-4% were obtained by G2P cleaning in the third system (*units_v1x_G2P*).

	MONO				
Lang	phon v1	grap v1	units v1	units v1x	units v1x G2P
CZ	61.7	61.0	42.0	51.6	54.8
EN	53.6	31.9	15.6	24.9	26.9
GE	43.2	37.0	20.0	27.1	28.6
PO	46.0	42.3	29.5	38.5	39.5
SP	57.2	55.3	39.0	47.6	49.2
RU	34.2	31.8	12.4	22.6	25.0
TU	70.1	70.7	57.5	61.2	63.1
VN	68.2	64.7	29.7	53.3	57.6

Table 2: Accuracy of MONO system with various dictionaries.

Table 3 shows results for advanced triphone system with speaker normalization. There degradation in comparison to baseline is smaller than in the monophone system. The context of triphones overall helped in word accuracy and difference according to baseline is around 3-8%.

	TRI2c (CMN)				
Lang	phon v1	grap v1	units v1	units v1x	units v1x G2P
CZ	68.1	68.4	51.1	60.7	63.1
EN	67.4	62.9	43.4	52.5	55.7
GE	58.4	57.7	40.9	47.5	51.3
PO	55.6	54.7	42.0	49.8	50.9
SP	68.7	68.7	52.2	60.5	63.9
RU	45.5	44.3	27.8	34.6	37.4
TU	75.3	75.7	66.0	70.4	71.8
VN	75.9	76.9	56.1	69.5	73.6

Table 3: Accuracy of TRI2c system (including CMN) with various dictionaries.

Last table 4 shows results of Sub-space GMM (SGMM) system for all variants of dictionaries. We can see another increase of accuracy for all systems. Deterioration between baseline (*phon_v1*) and best unit system (*unit_v1x_GP2*) is 2-7% depending on the language.

	SGMM (22k)				
Lang	phon v1	grap v1	units v1	units v1x	units v1x G2P
CZ	68.9	69.0	53.3	61.8	64.1
EN	70.5	65.7	47.2	56.5	59.1
GE	61.7	61.8	44.7	51.6	55.2
PO	57.3	57.1	42.4	51.3	54.2
SP	71.0	71.2	55.1	63.5	66.5
RU	48.1	47.1	29.4	37.3	40.8
TU	76.8	77.8	68.2	72.3	74.3
VN	80.4	81.7	63.3	74.3	78.5

Table 4: Accuracy of SGMM system with various dictionaries.

Presented results of generated dictionaries shows, that diarization technique is suitable and applicable in task of creating dictionaries for unseen or low-resource languages. The results of unit systems do not reach the accuracy of baseline system, but the technique could be further improved, e.g. by using features obtained from SGMMs or by using iVectors. Also the phase of clustering and labeling could be upgraded.

5 CONCLUSION

Diarization technique, proved to be appropriate approach for automatic generation of pronunciation dictionaries and is alternative to grapheme-based dictionaries. Combination of SGMM modeling with dictionaries obtained by diarization seems to be the right way, how to produce LVCSRs, for variety of low-resource 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] S. Besling. Heuristical and statistical methods for grapheme-to-phoneme conversion. In *Proceedings of* KONVENS-*94*, 1994.

[2] A.W. Black, K. Lenzo, and V. Pagel. Issues in building general letter to sound rules. pages 77–80, 1998.

[3] T. Fukada, T. Yoshimura, and Y. Sagisaka. Automatic generation of multiple pronunciations based on neural networks and language statistics. *Speech Communication*, 27:63–73, 1999.

[4] M. Huijbregts, M. Mclaren, and D. van Leeuwen. Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4436 –4439, may 2011.

[5] M. Janda, M. Karafiát, and J. Černocký. Dealing with Numbers in Grapheme-Based Speech Recognition. In *Proceedings of 15th International Conference on Text, Speech and Dialogue*, volume 2012 of *Lecture Notes in Computer Science, 2012, Volume 7499*, pages 438–445. Springer Verlag, 2012.

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