ARCHIVES OF ACOUSTICS
28, 4, 293–303 (2003)

LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM FOR POLISH

K. MARASEK

Polish–Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warszawa, Poland
kmarasek@pjwstk.edu.pl

This paper describes the development of the LVCSR (Large Vocabulary Continuous Speech Recognition) system for Polish. All system components have been built from scratch: grapheme-to-phoneme converter, language models and acoustic models. Test results for twenty thousands word vocabulary continuous speech recognition (read sentences) are given. The system can be used as a basis for application oriented continuous speech recognition.

1. Introduction

This paper describes the development of the LVCSR (Large Vocabulary Continuous Speech Recognition) system for Polish.

Despite of fast progress in speech technology, Polish seems not to be in the mainstream of works (see [3, 13, 16] for recent Polish-related efforts in automatic speech recognition). Beside LVCSR systems for main Germanic or Romance languages numerous attempts for Slavic languages are known: e.g. for Czech [7], Slovenian [5] or Russian [10]. The most obvious reason for the delay in the developments for Polish language was a lack of speech and language resources available for the scientific community. The outcomes of recent state- and international project enable to catch-up the status of other languages. The small project initiated at PJIIT should close the gap.

The elements of the typical speech recognition system are given in Fig. 1.

Speech signal is converted to a sequence of spectral and temporal features. Acoustic models represent basic units of speech and, given the features, estimate the probability of the occurrence of the unit in the signal given. Language model control the allowed sequence of words and estimates the probability of the occurrence of the sequence of words. Finally, the search module tries to find the sequence of words that best matches the observed signal.

Speech recognition typically evolves statistical modelling of language and speech, thus the recognition system need to be trained on a sufficiently large database. For the

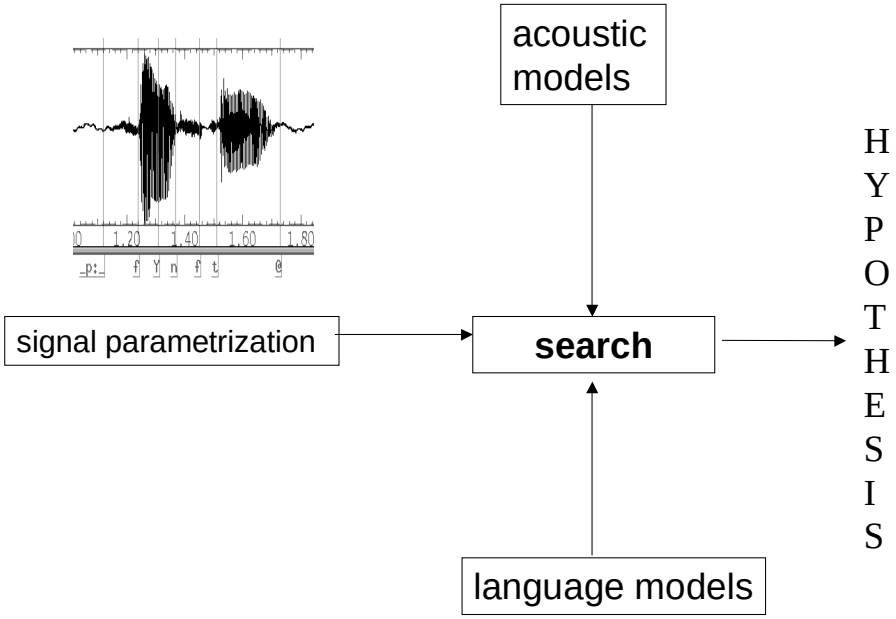

Fig. 1. Elements of speech recognition system.

acoustic models, depending on needs, the system could be speaker dependent (trained only for a given speaker) or speaker independent, enabling any speaker to use the system. For the second case, the size of training database grows significantly – repetitions of the same word (or phone) by different speakers are needed to balance the pronunciation variability.

The language model describes the domain of the recogniser. Typically, we could recognize isolated words (or short utterances) with relations described by the finite-state grammar (isolated word recognition) or connected speech, in which the probabilities of word sequences are given in a form of a stochastic language model.

The size of vocabulary could be related to the complexity of the recognition task. Typically, small vocabulary size describes systems with several hundreds of words, medium up to 2000, while large vocabulary systems are able to recognize several thousands of different words.

The system described below will recognize continuous speech, with 20 k (20 thousands) words vocabulary and is speaker independent.

One of the crucial system components is a dictionary, which contains the phonetic transcription of words (i.e. their phonetic equivalent). In the system presented the dictionary contains phones (fundamental acoustic units of speech), which are further expand to context dependent units.

The components necessary to establish the LVCSR system are depicted in Fig. 2 and described in details in the next sections.

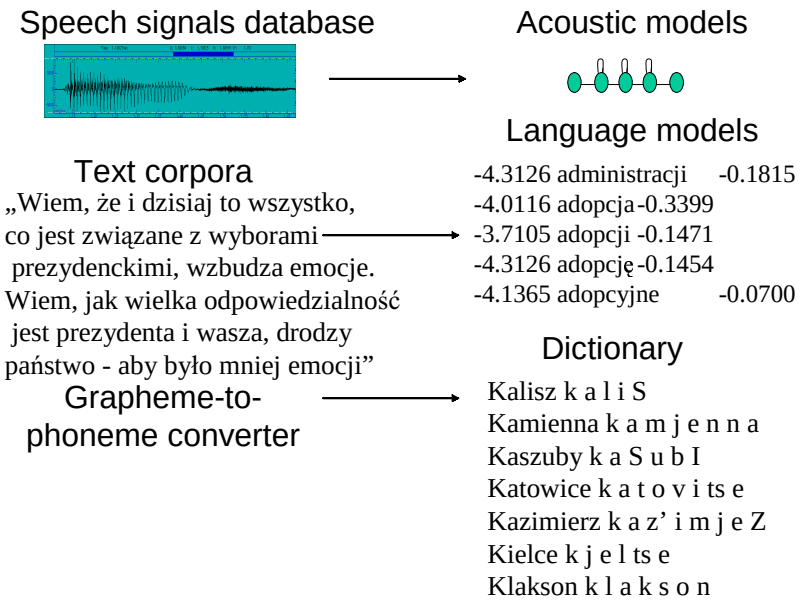

Fig. 2. LVCSR training components.

2. Speech signal databases

Speech data necessary for training the acoustic models have been collected from various sources. Mainly, the SpeeCon database has been used [11]. The database consists of 600 speakers (demographically, age and gender balanced) data including read and spontaneous passages and isolated words/commands recorded with 4-channels (close-talk, lavalier, desktop and far-field microphones) at various places (living-room, public place, child room, car). The close-talk channel data are transliterated and additional labels are added for non-speech noises:

- Speaker noises (breathing, lip smacks, etc.) and filler words (hesitations, uhms, etc.)
- External noises (stationary or not)

As the database is currently in preparation, only part of the data has been used for the acoustic modelling (ca. 200 speakers, no spontaneous speech recordings).

The second source of data was the WORDS database recorded at IFTR[(1)] , which contains isolated word recordings of 100 speakers (5 lists circulated over speakers, ca. 450 prompts per speaker with about 300 common phrases, ca. 1000 words vocabulary).

Finally, BABEL Polish speech database [1] has been used (originally recorded with 20 kHz sampling frequency, part of the material was phonetically transcribed and time-aligned, but the transcription wasn’t used in the training of the LVCSR system).

[(1)] Herewith I would like gratefully acknowledge the help of Prof. R. Gubrynowicz from Institute of Fundamental Technological Research, Polish Academy of Sciences.

Speech was picked-up with high quality desktop or close-talking microphones. All signals are stored in 16 kHz, 16-bit linear format.

3. Grapheme-to-phoneme converter

System uses a set of 37 phonemes to describe the basic sound patterns of Polish. The phone set, their SAMPA symbols and examples are given in Appendix[2] .

Converter accepts text coded in ISO-8859-2 lowercase.

The conversion is done automatically using machine learning approach. The rules have been learned from wide set of examples [4]. Alignment of letters and phones was achieved through introduction of empty phones (for the case that 2 letters represents only one phone) and combined phones (if one letter produces more than one phone), e.g. phonetic transcription of word *rząd* has following form: /_ Z o+n t/. Proper alignment of text and its phonetic transcription allows application of any classification methods to maximize the correctness of symbol conversion. Applied learning procedure was similar to the one used in the MBRDICO project [6]. Text is observed through a window of constant length (three graphemes to the left and three graphemes to the right serve as a context for a given character). Classification is performed by the decision tree (prepared using id3 program [8]). For each word only one pronunciation is generated. The decision tree is very small (7 kB), works fast with very high accuracy (typical error rate is smaller than 0.5%) for words of Polish origin. Foreign words have to be converted by hand[3] .

Post-processing rules are used to refine generated transcriptions of pronunciations (e.g. devoicing final consonants if the next word starts with unvoiced consonant).

4. Text corpora

Unfortunately, there is no public available annotated and sufficiently big corpus of written Polish. Thus, we decided to use non-normalized text data, prepare them and use to train the language model. The steps of text normalization are described below.

1. The raw text data (numbers written with digits, text containing abbreviations, etc.) contains speeches from Polish Parliament collected during 10 years[4] . While these are directly transcribed speeches, we hope to get more speech related language model. Small corpus (3700 sentences, ca. 39000 words) of newspaper texts was added to the main corpus to broad the scope of data. The whole corpus contains ca. 44 million words in more than 2800000 sentences. Not all the data have been used for language model training, mostly due to problems with proper text normalization.

---

[2] Appendix has been prepared with Prof. R. Gubrynowicz.

[3] Foreign words are those which pronunciation differ from Polish, e.g. *mail* is pronounced as /m e j l/ instead of /m a j l/.

[4] Herewith I would like to acknowledge help of Mr. Z. Jabłoński, head of the Polish Parliament's Computer Centre.

2. Text pre-processing splits the input into a sequence of tokens by separating the input where white space occur, deletes word-final punctuation marks, etc. Each unit (token) may contain single word, abbreviation, number, etc.

3. Sentence delimitation. Lists of tokens are linked into sentences. Consecutive sentences are then processed one by one. Marks for sentence begin and end are added.

4. Abbreviation recognition and expansion. This is done by look-up of external directory of abbreviations (actually about 500 abbreviations).

5. Numerals recognition and expansion. Numbers are converted to words using a grammar. Actually system correctly translated numbers from one billionth (0.00000001) up to one billion.

6. Correction of numerals depending on following word (numerals declension). The inflection endings of numbers depend on declination form of following word is found using the freely available Polish morphological analyser SAM-95 [12]. Script reads the incoming numerals and the following word. Depending on the grammatical form of the word the form of numeral is adjusted. Only first hypothesis delivered by the analyser was taken into account. Actual implementation of the numerals declension works well for integer numbers up to one billion. The accuracy of the numerals form correction depends on the accuracy of the analyser. The error rate has not been systematically analysed yet, but is less than 10%.

5. Language model

Normalized texts are used for language model preparation.

Statistical language model (LM) aims to represent the basic relation between short word sequences in the natural language. From the speech recognition point of view, we need a model, which generates all allowed word sequences for a given language. Current approach is to use a generative grammar or a stochastic language model. Grammars however, can be practically constructed only for very narrow domains, thus for LVCSR tasks stochastic language models are used [2].

Let $L = w_1^N = w_1, w_2, \ldots w_N$ be a word sequence and w_i's are the words that make up the hypothesis. The purpose of the language model is to calculate the probability $P(L)$, which can be computed using chain rule:

$$P(L) = \prod_{i=1}^{n} P\left(w_i | w_1^{i-1}\right), \tag{1}$$

where w_i^{i-1} is called history (h) or context of the word w_i. The commonly used simplification is to shorten the history to a n-gram LM (regardless of i) to $n-1$ words preceding the word:

$$h_i \approx w_{i-n+1} \ldots w_{i-1}. \tag{2}$$

This assumption leads to great reduction of the statistics needed to be collected to compute $P(L)$, however even then the number of parameters to be estimate is huge

(10^9 probabilities in case of tri-grams for 1000 words vocabulary). Another factor is the sparseness of real text data: most correct word sequences appear very rare even in very large text corpora [2]. The answer for that is use of smoothing techniques.

Language Modelling Toolkit [9] has been used to prepare the LM for the described text corpus. Bi-gram and tri-gram models are estimated using maximum likelihood method. Discounting used for estimation of zero-frequency cases is performed proportionally to a less specific context h': bi-gram distribution is used when tri-grams are estimated and uniform distributions when uni-grams are used. According to the backing-off scheme the n-gram probability is smoothed by selecting the best available approximation [2] of:

$$P(w|h) = \begin{cases} fr^*(w|h) & \text{if } fr^*(w|h) > 0, \\ \alpha_h \lambda(h) P(w|h') & \text{otherwise,} \end{cases} \tag{3}$$

where $fr^*(w|h)$ is a discounted conditional frequency, such that ($c(\,)$ denotes a number of occurrences) $0 \leq fr^*(w|h) \leq fr(w|h),\ fr(w|h) = c(hw)/c(c(h)$. The zero-frequency probability $\lambda(h) = 1.0 - \sum\limits_{w} fr^*(w|h)$ is redistributed over the words never observed in the context h, and a_k is normalization term assuring that $P(w|h)$ sums up to 1. For each n-gram *hw* the corrected frequency is computed. If the actual number of occurrences of n-gram *hw* is $c(hw)$, then the modified count is $d(c(hw))c(hw)$, where $d(c(hw))$ is called *discount ratio*. In the Good-Turing discounting used in the presented model, the discount ratio is equal to

$$d(hw) = (hw+1)n(hw+1)\,/hwn(hw), \tag{4}$$

where $n(hw)$ is the number of events which occur *hw* times. The discounting is only applied to counts which occur fewer than K times, where typically K is chosen to be around 7.

The quality of the language model is usually measured using perplexity, which describes how good the model can predict words in the text. For n-gram LM perplexity is computed as

$$PP = 2^{-\frac{1}{M}\sum\limits_{i=1}^{M} \log_2 \hat{P}(w_t|h_{t)}}. \tag{5}$$

For the model given, the bi-gram perplexity ranges from 54.86 for 3 k vocabulary to 74.41 for 64 k tested on 1000 randomly selected sentences which is in line with the results obtained for other languages [2, p. 205].

Better estimations of word probabilities could be obtained by clustering words into classes. That seems to be a preferred solution for Polish, due to rich declension and conjugation. Works on that are ongoing and results reported in [15] are very promising.

6. Acoustic models

3-states, left-to-right hidden Markov models are used to model the acoustics of speech. Each phone is modelled within the acoustic context of adjacent phones (so called triphones) including cross-word boundaries, with exceptions for silence and four additional noises (stationary, intermittent, speaker and filler noise) modelled without context (monophones). Feature vector is extracted every 10 ms from 16 kHz, 16-bits data (pre-emphasis coefficient $\alpha = 0.97$). Observation window of 25 ms is used. 24 triangular power Mel filter coefficients are computed for every window of pre-emphasized speech using FFT binning in the 80 to 7500 Hz frequency range. These coefficients are converted to cepstrals, using cepstral liftering (of range 22) to rescale them so that they have similar magnitudes [17]. Additionally, delta and delta-delta of window parameters are computed (change and acceleration of MFCC parameters between following windows). Thus, the full feature vector comprises of 38 parameters (12 cepstral coefficients, 12 delta-cepstrals, 12 delta-deltas, delta of energy and delta-delta energy).

Baum–Welch re-estimation procedure has been used. Continuous probability density functions (pdfs) are modelled using mixture of 3 Gaussians (normal pdf's distributions).

State clustering has been done using both data driven procedure and decision tree, asking questions about the acoustic/articulatory context of the model states. This was done to ensure that all state distributions could be robustly estimated. Resulting set of models uses 1069 states (clustering likelihood increase ratio=2000, see [17] for procedure details).

The overall amount of speech data used in training is about 100 hours of recordings (including silence).

7. Results

The acoustic models and language models have been tested on read speech: 360 sentences spoken by 12 speakers (part of the SpeeCondatabase, recordings of those speakers were unseen during training). The dictionary includes 20000 words. Results are summarized in Table 1. Results[(5)] are given on a sentence level (line started with SENT), word level (line started with WORD) and individually for all speakers (lines started with speaker codes 223–238). The sentence correctness is measured as a percentage of sentences for which all words in a sentence are correctly recognized (33.24%). Additionally, following statistics are given: the number of correct labels (H), the number of deletions (D), the number of substitutions (S), the number of insertions (I), N – the total number of labels in the defining transcription files (N) and M – number of mispronunciations.

Thus, the correctness is defined as

$$\mathrm{Corr} = \frac{H}{N} 100\% \tag{6}$$

[(5)] Computed using HResults program from the HTK suite [17].

and accuracy is computed as

$$\mathrm{Acc} = \frac{H - I}{N} 100\%. \tag{7}$$

Table 1. Recognition results for Polish sentences (20 k vocabulary, 12 speakers, 358 sentences).

Speaker Results	
spkr(sex): % Corr(% Acc)	[Hits, Dels, Subs, Ins, # Words] % S.Corr [# Sent] Mis
223(F): 87.11(86.84)	$[H = 331,\ D = 27,\ S = 22,\ I = 1,\ N = 380]\ 26.67\ [N = 30]\ M = 2$
224(F): 86.09(84.78)	$[H = 328,\ D = 33,\ S = 20,\ I = 5,\ N = 381]\ 30.00\ [N = 30]\ M = 3$
225(M): 77.69(75.85)	$[H = 296,\ D = 48,\ S = 37,\ I = 7,\ N = 381]\ 26.67\ [N = 30]\ M = 5$
226(M): 89.68(89.40)	$[H = 313,\ D = 21,\ S = 15,\ I = 1,\ N = 349]\ 41.38\ [N = 29]\ M = 1$
227(F): 89.92(89.65)	$[H = 330,\ D = 23,\ S = 14,\ I = 1,\ N = 367]\ 33.33\ [N = 30]\ M = 2$
228(F): 88.83(88.83)	$[H = 326,\ D = 24,\ S = 17,\ I = 0,\ N = 367]\ 33.33\ [N = 30]\ M = 2$
229(M): 87.43(87.16)	$[H = 320,\ D = 30,\ S = 16,\ I = 1,\ N = 366]\ 26.67\ [N = 30]\ M = 7$
230(F): 89.94(89.94)	$[H = 322,\ D = 25,\ S = 11,\ I = 0,\ N = 358]\ 24.14\ [N = 29]\ M = 10$
231(F): 83.38(82.27)	$[H = 301,\ D = 38,\ S = 22,\ I = 4,\ N = 361]\ 36.67\ [N = 30]\ M = 6$
232(M): 89.72(88.61)	$[H = 323,\ D = 14,\ S = 23,\ I = 4,\ N = 360]\ 46.67\ [N = 30]\ M = 0$
233(F): 90.56(90.00)	$[H = 326,\ D = 14,\ S = 20,\ I = 2,\ N = 360]\ 46.67\ [N = 30]\ M = 2$
238(M): 86.47(85.68)	$[H = 326,\ D = 26,\ S = 25,\ I = 3,\ N = 377]\ 27.59\ [N = 29]\ M = 3$
Overall Results	
SENT: % Correct=33.24 $[H = 119,\ S = 239,\ N = 358]$	
WORD: % $\mathrm{Corr} = 87.17$, $\mathrm{Acc} = 86.51$ $[H = 3852,\ D = 324,\ S = 243,\ I = 29,\ N = 4419]$	

Generally, the word recognition rate is high, while the sentence correctness is a bit unsatisfactory. The results can be easily explained by the fact, that the average sentence length was about 11 words and for almost all speakers mispronunciations have been observed.

Recordings are done in quite noisy environments (noise level ranges from 42 to 60 dBA with average over 50 dBA).

8. Conclusions

The paper reports the preparation of the LVCSR system for Polish. All system components have been built from scratch: grapheme-to-converter, language model and acoustic models. The preliminary test results for 20 k vocabulary are very promising, however a lot of improvements can be done to improve recognition accuracy:

• the general language model used in the experiments could be limited to the certain application domain,

• a class language model should be prepared, as noted in [15],

• gender specific acoustic models and VTLN (vocal tract length normalization) transformation.

The presented system will be used as a test-bet for further research on Polish speech recognition systems.

Appendix

The official set of SAMPA symbols for Polish can be found under: http://www.phon.ucl.ac.uk/home/sampa/Polish.htm. However, some modifications were introduced to mark in more clear way their phonological importance, e.g. for nasal vowels e∼ and o∼ whose existence in Polish is questioned by some authors.

Table 2.

Consonants		
Symbol	Word	Transcription
PLOSIVES		
p	pat	pat
b	bat	bat
t	test	test
d	dym	dIm
k	kat	kat
g	gen	gen
AFFRICATES		
ts	coś	tsos'
dz	dzwon	dzvon
ts'	ćwicz	ts'fitS
dz'	dźwięk	dz'vje∼k
tS	czyn	tSIn
dZ	dżin	dZIn
FRICATIVES		
f	fin	fin
v	waga	vaga
s	syk	sIk
z	zez	zes
S	szyk	SIk
Z	żyto	ZIto
s'	świt	s'fit
z'	źle	z'le
x	hak	hak

Table 2 [cont.]

Consonants		
Symbol	Word	Transcription
NASALS		
m	mak	mak
n	nasz	naS
n’	koń	kon’
N	gong	goNg
LATERAL		
l	luk	luk
APPROXIMANTS		
r	rak	rak
w	łuk	wuk
j	jak	jak
VOWELS		
i	tik	tik
I	typ	tIp
e	test	test
a	pat	pat
o	pot	pot
u	puk	puk
e∼	tę	te∼
o∼	tą	to∼

References

[1] BABEL, Copernicus # 1304 project, P. ROACH *et al.*, BABEL: An Eastern European Multi-Language Database, ICSLP-96, pp. 1982–1986, Philadelphia.

[2] R. DE MORI [Ed.], *Spoken dialogues with computers*, Academic Press, 1998.

[3] S. GROCHOLEWSKI, *The use of HMMs for modelling Polish triphones* [in Polish], Speech and Language Technology, W. JASSEM [Ed.], **5**, 59–76, 2001.

[4] L. MADELSKA, M. WITASZEK-SAMBORSKA, *Phonetic transcription. A set of exercises* [in Polish], Wydawnictwa UAM, Poznań 1997.

[5] S. MARTINČIĆ-IPŠIĆ, J. ŽIBERT, I. IPŠIĆ, F. MIHELIČ, *Speech recognition of Slovenian and Croatian weather forecasts*, Proceedings B of the 5-th International Multi-Conference Language Technologies, 14–15-th October 2002, Ljubljana, ISBN 961-6303-42-2.

[6] MBRDICO project, http://tcts.fpms.ac.be/synthesis/mbrdico.

[7] J. PSUTKA, *et al.*, *Automatic transcription of Czech language oral history in the MALACH project: resources and initial experiments*, Text, Speech, and Dialog Workshop, TSD2002, Brno 2002.

[8] J.R. QUINLAN, *Introduction of decision trees*, Machine Learning, **1**, 81–106, (1986).

[9] R. ROSENFELD, *Adaptive statistical language modelling: a maximum entropy approach*, PhD Dissertation, CMU 1994.

[10] T. SCHULTZ, M. WESTPHAL, A. WAIBEL, *The global phone project: Multilingual LVCSR with JANUS-3*, Multilingual Information Retrieval Dialogs: 2-nd SQEL Workshop, pp. 20–27, Plzen, April 1997.

[11] SpeeCon IST-1999-10003 EC project, http://www.speecon.com.

[12] K. SZAFRAN, *The morphological analyser SAM-95* [in Polish], ftp://mimuw.edu.pl.

[13] A. WIŚNIEWSKI, *Automatic speech recognition on the basis of hidden Markov models – problems and methods* [in Polish], Biuletyn Instytutu Automatyki i Robotyki WAT, **12**, 3–83, (2000).

[14] J. WELLS, *SAMPA – (Speech assessment methods phonetic alphabet)*, http://www.phon.ucl.ac.uk /home/sampa/home.htm.

[15] E. WHITTAKER, P. WOODLAND, *Comparison of language modelling techniques for russian and english*, Proceedings of ICSLP'98, Sydney 1998.

[16] A. WRZOSKOWICZ, *Hidden Markov models for noisy speech recognition*, Proc. Eurospeech'93, pp. 1591–1594.

[17] S. YOUNG, *et al.*, The HTK Book, http://htk.eng.cam.ac.uk.