ARCHIVES OF ACOUSTICS
32, 1, 119–126 (2007)

POLISH LVCSR IN THE JANUS SYSTEM

Preliminary results for the SpeeCon database

Krzysztof MARASEK

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

(*received October 16, 2006; accepted December 16, 2006*)

This paper describes the development of the LVCSR (Large Vocabulary Continuous Speech Recognition) system for Polish, using the Janus system developed at the University Karlsruhe/Carnegie Mellon University. The system has been tested on the selected material from the SpeeCon database. Test results for sentences read by 16 speakers are given. The system shows good performance and can be used as a basis for further development of modern speech recognition technology for Polish.

Keywords: speech recognition, LVCSR, Janus JRTk.

1. Introduction

Automatic speech recognition (ASR) has been a research topic for many years already and its commercial applications are used almost everywhere, starting from command and control for selected devices and ending at sophisticated dialogue systems for customer support over telephone line. Despite the technology progress, the deployment of the state-of-the-art techniques to the ASR of Polish is substantially delayed. There are several reasons for that: limited interest of Polish scientific institutions, no commercial interest at big companies for the Polish language, lack of resources needed for preparation of the ASR systems (spoken and written language corpora) and limited availability of modern software (most of the ASR systems are commercial). Thanks to the EU-projects and new Polish projects, more data become available to study the peculiarities of Polish spoken and written language.

1.1. Automatic speech recognition

Generally, in the statistical framework of the ASR, the task is to find such a sequence of words W for which the conditional probability $P(W|A)$ is maximized, where A is an

observed sequence of acoustic features describing incoming speech signal. Using the Bayes rule, this can be rewritten as:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)},$$

where $P(A|W)$ is the *a priori* probability of observing the feature sequence A for a given sequence of words W, the $P(W)$ is the probability of occurrence of a given sequence of words W, and $P(A)$ is probability of acoustic sequence A to be observed. Thus, $P(A|W)$ defines the acoustic model, while $P(W)$ is the language model. Due to continuous nature of the parameter space $P(A)$ is not easy to estimate. For given acoustic conditions it can be omitted without loss of generality, because we are indeed interested in finding the best fitting word sequence W:

$$W = \arg\max_W \frac{P(A|W)P(W)}{P(A)} = \arg\max_W P(A|W)P(W).$$

Acoustic modelling is usually performed using Hidden Markov Models (HMMs) and the training of models involves estimation of probability density functions with certain optimisation criteria (forward-backward training, maximum mutual information, discriminative training, etc. [3]). The language model may have several forms depending on the recognition task – it can be a simple context-free grammar for a limited set of recognized utterances, or, in the case of continuous speech, it may have a form of a statistical language model, where probabilities of N-words tuples are defined.

The paper describes initial experiences of using the Janus speech-to-speech translation system. In the first part, the JRTk-Janus Toolkit is briefly described. In the second part, data used for acoustic modelling are presented. The third part describes acoustic modelling and front-end of the recognizer. Next the language models used in experiments are presented. Finally, the results of the recognizer evaluation are given.

1.2. Janus recognition toolkit JRTk

The JANUS is a speech-to-speech translation system developed by the Interactive Systems Labs at the Karlsruhe University and Carnegie Mellon University. The JRTk [2], being a part of JANUS, is a flexible Tcl/Tk script-based environment. It is implemented in an object-oriented and portable paradigm allowing work on most hardware/software platforms (Windows, Linux, Solaris). The methods implemented in JRTk enable researchers to build state-of-the-art speech recognizers and allows them to develop, implement, and evaluate new methods. The Tcl/Tk shell allows access to core objects and relatively easy manipulation of data, also using a set of predefined scripts.

In the JRTk Version 5, the IBIS decoder can be used. The IBIS decoder is a one-pass decoder that is based on a re-entrant single pronunciation prefix tree and makes use of the concept of linguistic context polymorphism [5]. This enables decoding in one pass, incorporating full linguistic knowledge at an early stage of the decoding process, using the same engine in combination with a statistical n-gram language model

as well as context-free grammars. It is also possible to use the decoder to rescore the lattices in a very efficient way. This results in a speed-up compared to the decoder in previous versions of the JRTk, which needed three passes to incorporate full linguistic knowledge.

JRTk includes also sophisticated training procedures (like speaker-adaptive training [4]) and advanced feature space modifications (LDA transformation, semi-tied full covariances, vocal tract length normalization [3]), as well as decoding speed-ups (bucket-box intersection [6]), among many others.

1.3. SpeeCon database

For Polish, a relatively small amount of speech databases exist. Indeed, so far only 3 databases meet international standards: BABEL (Copernicus #1304 project) which, however, never have been officially released, SpeechDAT-East (telephone speech, INCO-Copernicus-977017) and SpeeCon (IST-1999-10003). Only the last one contains the speech material suitable for preparation of speaker-independent large vocabulary continuous speech recognition system (LVCSR).

The main target of the SpeCon project was collection of the speech data for at least 20 languages (and regional dialects), including most of the languages spoken in Europe. Polish recordings have been done at the PJIIT under the contract with Sony.

The database consists of 550 adults' recordings and 50 children recordings. Recordings were done in several real recording environments, defined as follows:

a. Office – is a relatively small room equipped with desks, usually or possibly with a computer. Noise level $L_{\mathrm{eq}} = 30 - 60$ dBA.
b. Entertainment environment – it is a living room with some furniture, places where people may sit down. A TV set or other audio sources may be turned on. $L_{\mathrm{eq}} = 30 - 65$ dBA.
c. Public Place – it is a very large room (hall) or open-air. A hall should have at least 3 walls and a ceiling over it; should be crowded and not too quiet. An open place has no walls and no closed ceiling (e.g. swimming pool). $L_{\mathrm{eq}} = 45 - 90$ dBA.
d. Car – for 4 or 5 passengers, in motion or not, with additionally specified speed and driving region (city, country, highway, etc.), $L_{\mathrm{eq}} = 28 - 80$ dBA.

The database is composed of read and spontaneous items.

The read part contains:

- phonetically compact sentences, 30 per speaker,
- phonetically rich words, 5 per speaker,
- general purpose words and phrases, 30 per speaker,
- application specific words and phrases, 220 per speaker.

The spontaneous part is composed of:

- free spontaneous items, ca. 15 per speaker,
- elicited answers, 17 per speaker.

Ca. 320 prompts per speaker were recorded. For each recording condition additional measurements of noise level and room impulse response were done.

Four microphone channels are recorded simultaneously: close-talk, lavalier, desktop and far-field (in a car, two special microphones have been used instead of desktop and far-field).

Beside orthographic transcription of the speech signal, the most important acoustic (non-speech) events present in the corresponding waveform files have been additionally marked. The symbol set of 7 markers is applied to indicate the presence and coarse categorization of the main speech distortions and acoustic events. This enables to keep as much speech in the database as possible, avoiding the need for taking recordings out from the corpus through some extra noises, disfluences or other events.

The lexicon for the recordings has been build up by various methods, including hand-annotation and generation by rule [8] with subsequent manual check. SAMPA phone symbols were used [8]. The pronunciation lexicon contains, for all transcribed words, number of their occurrences and the list of their phonemic representations. All truncated or mispronounced words and non-speech events are not included. Multiple transcriptions are supported in case of foreign words and spelling. The lexicon for adult speakers contains ca. 30000, and for children about 4000 items. More details on annotation and collection of SpeeCon data are given in [9]. The whole database needs approximately 140 GB.

2. Acoustic modelling

The HMM models have been trained using the close talk data. Recordings of 400 speakers have been selected from the database and only those recordings are taken which do not contain any mispronunciations nor any truncated words. Out of that a JRTk database was built, enabling an easy access to all utterances.

The standard topology of the HMM models was set – 3-states, left-to-right models, with emission probabilities estimated by mixture of Gaussians. Monophones have been trained in the first run, which in the second run have been expanded to polyphones for better context-dependent modelling. The phone set comprises 37 basic Polish phonemes [8] as defined in SAMPA set, 4 additional noise-dependent units (#sta# for stationary noise, #int# for intermittent noise, and #fil# for hesitations, #spk# for speaker noises such like lip smacks, especially important for close talk recordings) and a silence model.

The parameterization of speech signals includes MFCCs (13 coefficients), their first and second derivatives computed every 20 ms. Cepstral mean subtraction and Vocal Tract Length Normalization [3] is done in speaker-dependent manner and the LDA transformation is performed for better orthogonalization of the feature space.

Roughly speaking, the typical development cycle of a system for new language consists of the following steps [10]:

1. Writing labels (time-alignments) on the new data with an existing system.

The models have been built from scratch. The first step was to use an existing English recognition system to make initial labelling of the data (necessary for forward-backward training of initial models). The rewriting rules prepare an initial mapping of English to Polish phonemes and the first labels.

2. Training a new context-independent (CI) system on these labels and re-writing labels with the newly trained system several times.

First, means and parameters of the LDA transformation matrix with context-independent phones as classes are computed. Then, samples of the phones for all the training data are collected and clustered, using k-means algorithm – initial models for the phones were formed and trained with four iterations of label training.

3. Generating context questions on the CI system, training the resulting polyphone system and computing a clustering tree.

The performance of ASR systems can be greatly improved by looking at phonemes in their various contexts. If an unspecified neighbourhood is considered, the term polyphone is used. Sub-polyphones are modelled depending on their left and right neighbour phonemes. To compromise between generalization and model accuracy, the clustering algorithms merge context-dependent models together. The clustering procedure is based on a decision tree and clusters states of the same phoneme in different contexts. The questions that split the branches of the decision tree are linguistically motivated and have been prepared manually based on the phonetic knowledge. The question is selected if the entropy loss is maximized by a split of a node into two child nodes when applying this question. The clustering procedure begins at a node which collects all polyphones belonging to a phoneme, thus a phoneme in its various contexts. From the previously defined question set, the question maximizing entropy loss by performing a split is selected. The clustering procedure is finished if the number of leaves reaches the upper limit or if there is not enough of a training material to create and train a new node. A reasonable context neighbourhood width depends on the language and available training material.

4. Train a full context-dependent (CD) system using the clustering tree and the labels from the CI system.

The final training step involves the same steps as those needed to build the CI models: new description files (for codebooks and distributions) using the results from a polyphone training are created, LDA matrix is found, samples of polyphones are collected, k-means clustering is done and initialised models are then trained using the EM algorithm.

3. Language modelling

In speech recognition, we need a model which generates all the allowed word sequences for a given language; for the LVCSR tasks, stochastic language models are common [1]. Let $L = w_1^N = w_1, w_2, ..w_N$ be a word sequence of w_i words. The purpose of the language model is to calculate the probability $P(L)$, which can be computed using the chain rule:

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

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:

$$h_i \approx w_{i-n+1}...w_{i-1}. \quad (1)$$

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 estimated is huge (10^9 probabilities in case of tri-grams for 1000 words vocabulary). Another problem is the sparseness of real text data: most correct word sequences appear very rare, even in a very large text corpora [2]. The answer to that is to application of smoothing techniques [8].

In the described experiments bi- and trigram models were used. They were trained using corpora of the speeches from the Polish Parliament collected over 10 years [8]. The whole corpus contains ca. 44 million words in over 2.8 million sentences. Not all the data were used for language model training, mostly due to problems with proper text normalization. The bi-gram perplexity ranges from 54.86 for 3 k vocabulary to 74.41 for 64 k vocabulary tested on 1000 randomly selected sentences.

The language model prepared on the SpeeCon sentences (set of 3000 sentences) was also used in tests. SRI Toolkit [11] was used for the language modelling with standard settings (trigrams with Good-Turing discounting and Katz back-off for smoothing).

4. Test results

The acoustic models were tested in the same conditions as those done in [8]. The test set contains 433 sentences spoken by 16 speakers, close-talk channel (part of the SpeeCon database, recordings unseen during training) in office and entertainment recoding scenarios. Full SpeeCon dictionary was used. Results are summarized in Table 1. Results (computed using HResults from HTK Toolkit) 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 (55.66%). Additionally, the 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), the total number of labels in the defining transcription files (N) and number of mispronunciations (M).

The results presented in Table 1 were obtained after lattice re-scoring for the best language model weighting.

Additionally, an experiment for studio recordings was done with desktop microphone recordings (conditions and sentences unseen in the training). Recordings of 16 speakers containing 900 utterances (with repetitions) of ca. 5 k words vocabulary were used. The dictionary file was built fully automatically with automated grapheme-to-phoneme conversion tool [8]. In the recognition results, the correctness on the sentence

level achieved 32.91% while on the word level 88.10% (ranging from 67.90% to 94.75% – the recordings have been done without any supervision, thus some words in the recordings were mispronounced or spoken with hesitation).

The decoding procedure was very fast. The new IBIS decoder works very fast, being several times faster than the HVite (HTK Toolkit) applied in previous experiments [8]. The standard HTK Viterbi decoder uses a precompiled static recognition network, which does not support long span language models and is complex for cross-word triphones [12]. IBIS decoder uses a tree-structured lexicon, tree-structured representation of models and fast approximation of probability of Gaussian mixtures [6], what greatly enhances the speed of the decoding process.

Table 1. Recognition results for Polish sentences (30 k vocabulary, 16 speakers, 433 sentences).

Speaker Results									
spkr:	%Corr	(%Acc)	[Hits,	Dels,	Subs,	Ins,	#Words]	%S.Corr	[#Sent]
223:	87.37	(85.79)	[$H = 332$,	$D = 10$,	$S = 38$,	$I = 6$,	$N = 380$]	20.00	[$N = 30$]
224:	96.38	(93.31)	[$H = 346$,	$D = 3$,	$S = 10$,	$I = 11$,	$N = 359$]	53.57	[$N = 28$]
225:	95.95	(94.70)	[$H = 308$,	$D = 2$,	$S = 11$,	$I = 4$,	$N = 321$]	64.00	[$N = 25$]
226:	96.56	(96.28)	[$H = 337$,	$D = 1$,	$S = 11$,	$I = 1$,	$N = 349$]	89.66	[$N = 29$]
227:	95.64	(95.64)	[$H = 351$,	$D = 4$,	$S = 12$,	$I = 0$,	$N = 367$]	86.67	[$N = 30$]
228:	95.86	(95.56)	[$H = 324$,	$D = 2$,	$S = 12$,	$I = 1$,	$N = 338$]	78.57	[$N = 28$]
229:	97.94	(97.25)	[$H = 285$,	$D = 1$,	$S = 5$,	$I = 2$,	$N = 291$]	79.17	[$N = 24$]
230:	94.16	(91.23)	[$H = 290$,	$D = 2$,	$S = 16$,	$I = 9$,	$N = 308$]	44.00	[$N = 25$]
231:	89.35	(85.57)	[$H = 260$,	$D = 4$,	$S = 27$,	$I = 11$,	$N = 291$]	24.00	[$N = 25$]
232:	96.67	(95.83)	[$H = 348$,	$D = 3$,	$S = 9$,	$I = 3$,	$N = 360$]	70.00	[$N = 30$]
233:	96.37	(95.17)	[$H = 319$,	$D = 0$,	$S = 12$,	$I = 4$,	$N = 331$]	60.71	[$N = 28$]
234:	89.88	(86.50)	[$H = 293$,	$D = 7$,	$S = 26$,	$I = 11$,	$N = 326$]	18.52	[$N = 27$]
235:	93.42	(91.36)	[$H = 227$,	$D = 5$,	$S = 11$,	$I = 5$,	$N = 243$]	52.38	[$N = 21$]
236:	95.00	(93.95)	[$H = 361$,	$D = 3$,	$S = 16$,	$I = 4$,	$N = 380$]	72.41	[$N = 29$]
237:	89.97	(85.27)	[$H = 287$,	$D = 4$,	$S = 28$,	$I = 15$,	$N = 319$]	40.00	[$N = 25$]
238:	92.59	(89.95)	[$H = 350$,	$D = 5$,	$S = 23$,	$I = 10$,	$N = 378$]	31.03	[$N = 29$]
Overall Results									
SENT: %Correct = 55.66 [$H = 241$, $S = 192$, $N = 433$]									
WORD: %Corr = 93.95, Acc = 92.14 [$H = 5018$, $D = 56$, $S = 267$, $I = 97$, $N = 5341$]									

5. Conclusions

The paper reports an the ongoing work. The JRTk has proven to be a useful tool for building a speech recognition system for a new language from scratch. The first results were obtained in a relatively short time, using for the training of the acoustic models only standard methods and procedures. I am sure that application of the novel techniques, implemented already in the JRTk system, will enable further and significant improvement of the recognition results. We can also state that the SpeeCon database forms a valuable set of recordings for both training of the speech recognizers as well as for study of the nature of Polish spoken language [9].

References

[1] DE MORI, *Spoken dialogues with computers: Signal processing and its applications*, Academic Press, Hardcover, 1998, ISBN 0122090551.

[2] FINKE M., GEUTNER P., HILD H., KEMP T., RIES K., WESTPHAL M., *The Karlsruhe–VERBMOBIL speech recognition engine*, Proceedings of the IEEE ASSP Conf., vol. 1, pp. 83–86 Munich 1997.

[3] HUANG, ACERO, HON, *Spoken language processing*, Prentice Hall, 2001, ISBN 0130226165.

[4] MATSOUKAS S., SCHWARTZ R., JIN H., NGUYEN L., *Practical implementations of speaker-adaptive training, proceedings of the 1997 DARPA speech recognition workshop*, Chantilly, Virginia, USA, February 2–5, 1997.

[5] SOLTAU H., METZE F., FÜGEN CH., WAIBEL A., *A one pass-decoder based on polymorphic linguistic context assignment*, Proceedings of the ASRU Workshop, Madonna di Campiglio, Italy, pp. 214–217, 2001.

[6] FRITSCH J., ROGINA I., *The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture Gaussians*, Proceedings of ICASSP 96, Atlanta, Vol. 2, pp. 837–840, 1996.

[7] ROACH P. *et al.*, *Babel: An eastern european multi-language database, proceedings of ICSLP-96*, Philadelphia, pp. 1982–1986, 1996.

[8] MARASEK K., *Large vocabulary continuous speech recognition system for Polish*, Archives of Acoustics, **28,** 4, 293–303 (2003).

[9] MARASEK K., GUBRYNOWICZ R., *Mutlilevel annotation in SpeeCon Polish speech database*, IMTCI (International Workshop on Intelligent Media Technology for Communicative Inteligence), Warszawa, Lecture Notes in Computer Science, Springer, pp. 58–67, 2004.

[10] IslSystem Documentation, Example Training/Testing Setup for Use with JRTk, The Ibis-gang, 22. November 2002, v0.1, Interactive Systems Labs, University of Karlsruhe, Germany, 2002.

[11] STOLCKE A., *SRILM – An extensible language modeling toolkit*, Proc. of ICSLP 2002, Denver, Colorado, pp. 901–904, 2002.

[12] WOODLAND P. C., ODELL J. J., VALTCHEV V., YOUNG S. J., *Large vocabulary continuous speech recognition using HTK*, Proceedings ICASSP'94, Adelaide, pp. 125–128, 1994.