Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license

Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, Filip Jurčíček

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{korvas,oplatek,odusek,zilka,jurcicek}@ufal.mff.cuni.cz

Abstract

We present a dataset of telephone conversations in English and Czech, developed to train acoustic models for automatic speech recognition (ASR) in spoken dialogue systems (SDSs). The data comprise 45 hours of speech in English and over 18 hours in Czech. All audio data and a large part of transcriptions was collected using crowdsourcing; the rest was transcribed by hired transcribers. We release the data together with scripts for data pre-processing and building acoustic models using the HTK and Kaldi ASR toolkits. We publish the trained models described in this paper as well. The data are released under the CC-BY-SA 3.0 license, the scripts are licensed under Apache 2.0. In the paper, we report on the methodology of collecting the data, on the size and properties of the data, and on the scripts and their use. We verify the usability of the datasets by training and evaluating acoustic models using the presented data and scripts.

Keywords: Acoustic data, orthographic transcriptions, ASR scripts

1. Overview

Recorded and transcribed speech is an important language resource for building ASR models. In this paper, we describe two datasets of transcribed telephone speech, an English one and a Czech one.

Individuals who set to build a spoken dialogue system or another application that employs ASR currently have few options to obtain the ASR system, none of which we consider very plausible. One option is to buy a ready-made ASR system, such as Nuance Dragon.[1] However, this option is costly and comes with a license that is likely to be too restrictive for research purposes.[2]

Another option is exploiting a cloud-based ASR, such as Google ASR.[3] It may provide state-of-the-art quality for many tasks (Morbini et al., 2013) and can be used for free. However, its licensing conditions are not clear, customisation to a task at hand is limited, and the service is not officially supported.

The third option is to build a custom ASR model using one of the ASR toolkits available. This allows domain adaptation and maximum customisability, especially with open-source toolkits such as Kaldi (Povey et al., 2011). The drawback is the need for a large enough amount of acoustic data and their transcriptions. Suitable acoustic datasets are available at least for English; however, they come with very restrictive licenses (Rousseau et al., 2012) and some of them are also costly, at least for non-LDC-members, e.g., CALLHOME American English Speech and Transcripts (Canavan et al., 1997; Kingsbury et al., 1997).

In this paper, we present *free* transcribed speech corpora for English and Czech bundled with working scripts for training ASR models, with the goal to foster research and applications using ASR. These resources will be particularly useful for developing SDSs which communicate with users over telephone, since this is the channel through which we collected the audio data. The data are released under the Creative Commons Share-alike (CC-BY-SA 3.0) license, the scripts under the Apache 2.0 license.

The paper is structured as follows. In Section 2, we describe how the data were collected and processed, and report the size of both datasets. In Section 3, we give an overview of the scripts we use to train ASR models, and in Section 4, we report on recognition results obtained with the models trained. We conclude by providing details on how the data and scripts can be obtained, in Section 5.

2. Data

The two datasets, the English one and the Czech one, have been collected in different ways. We shall describe the data collection procedures in the following.

2.1. Collecting the English data

The English recordings were collected from humans interacting via telephone calls with statistical dialogue systems, designed to provide the user with information on a suitable dining venue in the town. The data collection process was run through the Amazon Mechanical Turk (AMT)[4] crowd-

[1] http://www.nuance.com/dragon/index.htm

[2] NDEV Mobile Policies explicitly forbid its use in development of speech recognition software (http://nuancemobiledeveloper.com/public/index.php?task=policies).
Publishing benchmarking information is probably forbidden as well (http://www.nuance.com/products/dragon-medical-360-network-edition/eula/index.htm), ruling out the possibility to report on the quality of ASR used in your system.

[3] The API is accessible at https://www.google.com/speech-api/v1/recognize, and its use described in a blog post at http://mikepultz.com/2013/07/google-speech-api-full-duplex-php-version/.

[4] https://www.mturk.com

THE ADDRESS

WHAT'S THE ADDRESS PHONE NUMBER AND PRICE RANGE

LOOKING FOR A RESTAURANT SERVING ANY KIND OF FOOD AND IT SHOULD BE FREE

Table 1: Sample transcriptions from the English data

sourcing platform. As a consequence, most of the recordings were produced by speakers of American English.

AMT workers that chose to work on this microtask were given a phone number where our dialogue system was awaiting a call, and an agenda specifying what parameters the venue should have. The entire conversation was logged and the recordings of the human's utterances serve as the basis of the present dataset. We also employed AMT workers to transcribe the recordings, which lead to cheap transcriptions with basic markup. In order to ensure high quality of the transcriptions, we only allowed workers who had completed 500 HITs[5] with an acceptance rate of over 97% to work on them. In addition, we included one sentence in every dialogue for which we had a gold-standard transcription. We used the gold transcriptions to evaluate workers' reliability.

A sample of the English data is shown in Table 1.

2.2. Collecting the Czech data

The Czech recordings were collected in three ways:

1. using a free Call-a-Friend phone service (FRIEND)
2. using the Repeat-After-Me (RAM) speech data collecting process
3. from telephone interactions with the PublicTransportInfo (PTI) SDS.

In FRIEND, native Czech speakers were invited to make free calls using an automated service interconnecting two callers on demand. In return, they gave us consent to record the calls. In RAM, volunteers called a number where they were asked to repeat sentences synthesized by a TTS. In PTI, they can interact with an SDS on phone to find public transport connections. In all three cases, the callers agreed that their recordings may be used by third parties, even for commercial purposes, before they started using the service. We have anonymized the data so that utterances that contained personal information were excluded and phone numbers of the callers are not included in the data.

All Czech recordings in the data were transcribed by professional transcribers. The transcriptions are orthographic and capture several kinds of non-speech events as well as incompletely pronounced words and foreign words used in Czech discourse.

A sample of the Czech data is shown in Table 2.

JE TŘEBA V TOMTO MEDIÁLNÍM SVĚTĚ ŽÍT

TAK A JÁ MÁM VYBALANCOVÁNO

PROTOŽE PROTOŽE VLASTNĚ U NÁS V TOM BLA CHYTLA TA STUPAČKA JAK JSOU TY GARZONKY TENKRÁT VÍŠ

Table 2: Sample transcriptions from the Czech data

dataset	audio	# sents	# words
English			
train	41:30	47,463	178,110
dev	1:45	2,000	7,376
test	1:46	2,000	7,772
Czech			
train	15:25	22,567	126,333
dev	1:23	2,000	11,478
test	1:22	2,000	11,204

Table 3: Size of the data: length of the audio (hours:minutes), number of sentences (which is the same as the number of recordings), number of words in the transcriptions.

2.3. Cleaning the transcriptions

In both languages, the collected transcriptions were normalised and filtered. The normalisation comprised uppercasing, discarding punctuation, ad hoc spelling correction and some lesser changes of a technical character.

When filtering the English data, we looked for transcriptions containing non-existent words or special symbols and discarded them together with the corresponding recordings.

The Czech recordings were transcribed by a few professional transcribers, hence we were not suspicious about the existence of words present in the transcriptions. Even though some special symbols had their meaning in the transcription markup, such as denoting incompletely pronounced words, we decided to discard all transcriptions containing special symbols along with their recordings, since for those data, the content of the recording (a sequence of phones and other sounds) cannot be reliably reconstructed from the transcription.

2.4. Data characteristics

Our ultimate goal in ASR development is to build one general acoustic model per language, to simplify the deployment of our dialogue systems. Hence, we put the data from different sources (FRIEND, RAM, etc.) together, distinguishing them only by the language. This makes no harm to acoustic model training as the data were always recorded through the same channel (telephone). To make the datasets more coherent, we ensure that all the recordings are 16bit 16kHz audio WAV files.

For both English and Czech datasets, we select and fix development and test parts of the data, treating the remainder as a training set. We suggest that any later results obtained from these data are based on the same split, for sake of comparability. Table 3 lists the size of the datasets in terms of recorded audio length, number of sentences (or, equally, number of recordings) and number of transcribed words, after cleaning.

The careful reader will note that the words-per-second ra-

[5] HIT (human intelligence task) is a term for one microtask used by AMT.

tio is almost twice higher for the Czech dataset. We suppose this is because speech segments in the recorded audio data were delimited by human transcribers in case of the Czech recordings, and by an automatic voice activity detector in case of the English recordings. Therefore, the English recordings contain more silence or noise around the actual speech segments.

3. ASR acoustic modeling

We verified the usability of the collected acoustic data by training acoustic models and evaluating their performance. In our ASR-related research, we use the Kaldi toolkit (Povey et al., 2011). Hence, we have experimented more with Kaldi, and use the scripts for the HTK toolkit (Young et al., 2006) merely as a starting point.

We release scripts for both toolkits and for both English and Czech. Importantly, the scripts can be easily made to work with other datasets, and after the addition of necessary language-specific components, also with other languages. The input data are expected to be in a simple format of pairs of files, an `X.wav` file with the recording, and an `X.wav.trn` file with the transcription.

The scripts for both toolkits code the recordings into mel frequency cepstral coefficients (MFCCs), their Δ and $\Delta\Delta$ features.

In the following, we first list all language-specific parts of our training scripts and then briefly describe our training procedures for both ASR toolkits.

3.1. Language-specific components of the training scripts

The only language-specific components included in the training scripts are:

- list of phones the language uses
- an orthography-to-phonetics mapping
- phonetic questions (only needed with HTK scripts; see Section 3.2 for details).

The orthography-to-phonetics mapping can be either enumerated in a pronouncing dictionary, or implemented in a script.

For English data, we derive the phonetic transcriptions using the CMU pronouncing dictionary[6] (version 0.7a), extended with about 250 words to cover the vocabulary of the collected utterances.

For Czech, we use a set of regular expressions implementing Czech rules of pronunciation (Psutka et al., 2006). The code for phonetic transcription of Czech is made available as one of the scripts.

3.2. Training using HTK

As the tool's name, HTK (HMM Toolkit), suggests, elements of speech are modeled using Hidden Markov Models (HMMs). Each element of speech (phone, triphone or other sound) is modeled by a sequence of states which generate a Gaussian mixture (GM) distribution over audio features. These Gaussians always have a diagonal covariance matrix. Our training scripts for HTK are derived from scripts for training acoustic models by Vertanen (Vertanen, 2006), which in turn "closely follows the steps in the tutorial in the HTKBook" (Young et al., 2006). Vertanen's scripts, evaluated on ARPA CSR Benchmark Tests Corpora (Garofalo et al., 2007; Linguistic Data Consortium, 1994), achieved state-of-the-art results at the time they were published. We added several extensions to Vertanen's recipes and focus only on these extensions in the following description.

Single-phone HMMs are initialised from a flat start, and after three rounds of Expectation-maximisation (EM) training, rough monophone models get trained. We fix the silence model (Young et al., 2006, Section 3.2.2) and force-align the coded audio data to their phone-level transcriptions to disambiguate between pronunciation variants in word realisations. After another four iterations of EM, we arrive at the final simple monophone model.

When the monophone models have been trained, we expand the inventory of symbols from phones to triphones and train models for these. Triphones are clustered and a single set of parameters is learned for each triphone cluster (this is known as *parameter tying*). We perform the clustering using decision trees built from a set of phonetic questions and state occupancy statistics for the training data. *Phonetic questions* is a term used to describe phonetically motivated criteria for grouping similar triphones. The phonetic questions we use were designed to closely follow the guidelines in the HTK book (Young et al., 2006). Since the phonetic alphabets for Czech and English differ, two sets of phonetic questions were constructed, one for each language. With the parameters tied, the models are retrained in another four iterations of EM. Finally, we gradually increase the number of Gaussians in the mixture describing each HMM state until there are 18 Gaussians per state for regular phone HMMs and 36 Gaussians per state for the silence model.

After the training, we export HTK acoustic models in a format suitable for the Julius ASR decoder.[7]

3.3. Training using Kaldi

The Kaldi toolkit is based on Finite State Transducers (FSTs), so the internal representation differs from HTK. However, we followed the same steps for acoustic training to ensure comparability of the resulting acoustic models. We designed the training procedure so that it is very similar to the training procedure using HTK.

Our training scripts for Kaldi were inspired by scripts for VoxForge data written by Vassil Panayotov (Panayotov, 2012).

As with HTK, we model elements of speech using HMMs generating audio features through Gaussians with diagonal covariance matrices.

We train a monophone model from flat start using the MFCCs, Δ and $\Delta\Delta$ features. We force-align the audio data (in the form of feature vectors) to HMM states for phones in the corresponding transcriptions. The triphone model is

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

[7] `http://julius.sourceforge.jp/en_index.php`

trained using Viterbi training. We cluster the triphones using a phonetic decision tree, which takes into account the immediate left and right context of a phone. Questions for nodes of the phonetic tree are generated automatically (Povey et al., 2011). The model with tied parameters is retrained similarly as in HTK scripts.

So far, we described the estimation of generative models, but Kaldi also implements a variety of discriminative training methods, including:

- Maximum likelihood linear regression (Leggetter and Woodland, 1995)
- Discriminative training based on the criterion of minimum phone error (MPE) (Povey and Woodland, 2002)
- Boosted Maximum Mutual Information (BMMI) (Povey et al., 2008).

Discriminatively trained feature transformations available in Kaldi include:

- Linear Discriminative Analysis (LDA) (Haeb-Umbach and Ney, 1992)
- Maximum Likelihood Linear Transformation (MLLT) (Gopinath, 1998)
- Feature-space minimum phone error (fMPE) (Povey et al., 2005).

In our choice of discriminative training, we were limited to non-speaker-adaptive methods. We present models obtained by BMMI training with LDA and MLLT feature transformations. The triphone model with LDA and MLLT transformations was trained using alignments from basic triphone model. We force-align the audio data from the model with LDA+MLLT and then train the BMMI model. Note that BMMI needs a *language model* (LM) in order to compute the objective function. Here we use the bigram LM as described in Section 4.

4. Evaluation

We use the scripts described in Section 3 to train acoustic models with complexity as similar as possible. Namely, we set the total number of Gaussians in GMs in the models to about 41k and 27k for Czech and English models, respectively.

We evaluate the trained models on the test set for the respective language, using zerogram and bigram LMs. For the zerogram LM, we decided to include only words from the test set. The motivation for this is to evaluate solely the quality of the acoustic models without being affected by a language model or presence of out-of-vocabulary (OOV) words in the test set. The zerogram LM for Czech contains 3,150 words, while the one for English contains only 345 words.

With the bigram LMs, we simulate a more realistic setup where the goal is to achieve favourable recognition error rates. The bigram LMs are computed from the training data for the particular language. Basic properties of the bigram LMs are summarised in Table 4. One can clearly see that complexity of the Czech data is much larger when compared to English data. This is because the Czech data includes FRIEND and RAM data with unconstrained speech while the English data consists of conversations with a limited-domain dialogue system.

language	# words	perplexity	OOV
English	648	5.4	15
Czech	12,540	135.0	894

Table 4: Size, perplexity and number of OOV words of bigram models used for evaluation. Perplexity and OOV figures are valid for the test set.

language	method	zerogram	bigram
Czech	tri $\Delta + \Delta\Delta$	64.5	60.4
English	tri $\Delta + \Delta\Delta$	50.0	17.5

Table 5: Word error rates on test set obtained using HTK and either a zerogram or a bigram LM.

For both HTK and Kaldi decoding, the language model weights were manually tuned on the dev set.

4.1. HTK results

We report on the Word Error Rate (WER) obtained by decoding the test sets with HTK tools[8] in Table 5. Observing the results, one can see the WER for English data is much lower when compared to Czech data. This is partly due to larger training data set and partly due to less complex language models. Regarding the absolute WER values, the results are in line with WERs of similar tasks (Laroche et al., 2011), confirming the usability of the presented datasets.

4.2. Kaldi results

The WERs obtained by decoding the test sets with the Kaldi toolkit[9] are shown in Table 6. Please note that when performing BMMI discriminative training, the acoustic models are optimised for bigram LMs. Therefore, we do not report on BMMI results for decoding with the zerogram LM. The results suggest that Kaldi achieves similar WER compared to HTK when using standard generative training methods and bigram LMs. One can obtain a substantial decrease in WER by using more advanced discriminative training methods.

5. Availability

The data and scripts are made available through the LINDAT/CLARIN repository, `lindat.cz`. As already noted, the data are released under the Creative Commons (CC-BY-SA 3.0) license and the scripts under the Apache 2.0 license.

The data and scripts are distributed in three parts at the following URLs:

- Czech data: `http://hdl.handle.net/11858/00-097C-0000-0023-4670-6`

[8] Namely, using the program HVite for decoding.

[9] Namely, using the program gmm-latgen-faster for decoding.

language/method	zerogram	bigram
Czech		
tri $\Delta + \Delta\Delta$	69.3	53.8
tri LDA+MLLT	65.4	51.2
tri LDA+MLLT+BMMI	–	48.0
English		
tri $\Delta + \Delta\Delta$	41.1	17.5
tri LDA+MLLT	37.3	17.2
tri LDA+MLLT+BMMI	–	12.0

Table 6: Word error rates on test set obtained using Kaldi. The 'tri $\Delta + \Delta\Delta$' row shows results for a basic generative model with triphones which is comparable to the model trained using the HTK scripts.

- English data: `http://hdl.handle.net/11858/00-097C-0000-0023-4671-4`
- all scripts and trained models described in this paper: `http://hdl.handle.net/11858/00-097C-0000-0023-466F-C`

Acknowledgments

We are grateful to the authors of the script implementing Czech rules of pronunciation, Nino Peterek, Pavel Ircing, Oldřich Krůza, and other contributors for allowing us te could publish it as part of our scripts.

We also thank the anonymous reviewers for their helpful comments and suggestions.

This research was funded by the Ministry of Education, Youth and Sports of the Czech Republic under the grant agreement LK11221, by the core research funding of Charles University in Prague, and by SVV project number 260 104.

The language resources presented in this work are stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

6. References

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English speech. LDC Catalog No.: LDC97S42.

John Garofalo, David Graff, Doug Paul, and David Pallett. 2007. CSR-I (WSJ0) complete. LDC Catalog No.: LDC93S6A.

Ramesh A. Gopinath. 1998. Maximum likelihood modeling with Gaussian distributions for classification. In *Proc. ICASSP*, volume 2, pages 661–664. IEEE.

Reinhold Haeb-Umbach and Hermann Ney. 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. ICASSP*, volume 1, pages 13–16. IEEE.

Paul Kingsbury, Stephanie Strassel, Cynthia McLemore, and Robert McIntyre. 1997. CALLHOME American English transcripts. LDC Catalog No.: LDC97T14.

Romain Laroche, Ghislain Putois, Philippe Bretier, Martin Aranguren, Julia Velkovska, Helen Hastie, Simon Keizer, Kai Yu, Filip Jurčíček, Oliver Lemon, et al. 2011. D6.4: Final evaluation of CLASSiC TownInfo and appointment scheduling systems.

Christopher J. Leggetter and Philip C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

Linguistic Data Consortium. 1994. CSR-II (WSJ1) complete. LDC Catalog No.: LDC94S13A.

Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Doğan Can, Panayiotis G. Georgiou, Shrikanth S. Narayanan, Anton Leuski, and David Traum. 2013. Which ASR should I choose for my dialogue system? In *Proc. SIGDIAL*, August.

Vassil Panayotov. 2012. VoxForge scripts for Kaldi. `http://vpanayotov.blogspot.cz/2012/07/voxforge-scripts-for-kaldi.html`.

Daniel Povey and Philip C. Woodland. 2002. Minimum phone error and I-smoothing for improved discriminative training. In *Proc. ICASSP*, volume 1, pages 105–108. IEEE.

Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau, and Geoffrey Zweig. 2005. fMPE: Discriminatively trained features for speech recognition. In *Proc. ICASSP*, volume 1, pages 961–964. Philadelphia.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah. 2008. Boosted MMI for model and feature-space discriminative training. pages 4057–4060. IEEE, March.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Josef Psutka, Luděk Müller, Jindřich Matoušek, and Vlasta Radová. 2006. *Mluvíme s počítačem česky [Speaking Czech to the computer]*. Academia, Prague.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Keith Vertanen. 2006. Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory, University of Cambridge.

Steve J. Young, Gunnar Evermann, Mark J. F. Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil C. Woodland. 2006. *The HTK Book,* *version 3.4*. Cambridge University Engineering Department, Cambridge, UK.