Euronews: a multilingual benchmark for ASR and LID

Roberto Gretter

FBK - Via Sommarive, 18 - I-38123 POVO (TN), Italy
gretter@fbk.eu

Abstract

In this paper we present the first recognition experiments on a multilingual speech corpus, designed for Automatic Speech Recognition (ASR) and Language IDentification (LID) purposes. Data come from the portal Euronews and were acquired both from the Web and from TV. The corpus includes data in 10 languages (Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish). For each language, the corpus is composed of about 100 hours of speech for training (60 for Polish) and about 4 hours, manually transcribed, for testing. Training data include the audio, some reference text, the ASR output and their alignment. 10 baselines were prepared - one for each language - using only the training data, and performance are evaluated on a subset of the test data. Also a LID system was implemented, capable to recognize words belonging to different languages in a continuous stream. Part of the corpus is freely available, for research purposes only, within the multilingual ASR benchmark for IWSLT 2014.

Index Terms: multilingual, speech corpus, light supervision, ASR, language identification

1. Introduction

In the last few years, Web crawling and TV recording have become main data sources for research and progress in Automatic Speech Recognition (ASR). Monitoring the Web to look for new resources is a fundamental activity, in particular for what concerns manually transcribed audio/video data and parallel corpora. In the past we started to collect data, on a daily basis, from the portal Euronews[1], which broadcasts news in several languages. Then, in the last year we designed a collected a multilingual speech corpus [1] for ASR purposes and in this paper we describe the first ASR and Language IDentification (LID) experiments we made on it. Other multilingual resources were created and made available to the research community in the past, such as Europarl [2] for MT translation and GlobalPhone [3] for multilingual ASR.

Audio data can be used to train Acoustic Models (AMs) for a target language in a complete unsupervised way, as described in [4, 5], where AMs are bootstrapped from those of a well trained source language, without using any transcribed data in the target language. But when the audio data come along with some text, which could be a partial transcription of it, other approaches are possible which exploit the so called "lightly supervised" training [6] which uses transcriptions close, but not exact, to what is spoken.

AM training, as for it, has been largely investigated by many researchers, who proposed approaches based on the usage of language specific [7, 8], language universal [9, 10] and language adaptive [11, 10] acoustic models. In general, the training of language specific acoustic models represents the best practice to adopt when a sufficient amount of audio recordings (i.e. hundreds of hours) is available for a given language. On the contrary, when a reduced set of training data (i.e. tens of hours or less) is available for a language, two different approaches can be used: **(i)** cross-language bootstrap [7] of a *target* language's AM starting from that of a well trained *source* language. Bootstrap can be followed by training, or adaptation, using the available set of training data of the target language; **(ii)** training of a universal set of acoustic models using a mix of training data in many languages [9] [10] [11], also possibly followed by language dependent adaptation.

The multilingual Euronews corpus is composed of about 100 hours of training data and about 4 hours, manually transcribed, of test data, for each of the following 10 languages: Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish. These languages cover most of the languages that are addressed in the European project EU-BRIDGE[2]. This paper is organized as follows: Sections 2, 3 and 4 describe the corpus, Section 5 and 6 describe the first ASR and LID experiments we did on it, Section 7 reports about the availability of the corpus for the research community. Finally, in Section 8 we draw our conclusions and outline future work.

2. Euronews: sources and data description

Euronews provides multilingual data through two main channels: a satellite TV and a Web portal. We acquire data from both of them.

2.1. Data from TV

International news are acquired from the satellite TV channel in different languages, potentially attractive as a source of comparable speech data. At present news are transmitted in 13 different languages (new languages are added over time): Arabic, English, French, German, Greek, Hungarian, Italian, Persian, (in the past Polish), Portuguese, Russian, Spanish, Turkish and Ukrainian, each one over a different audio channel in the digital TV stream. As the video content is the same for all of the audio channels, the news are time aligned.

From ASR perspective, data cannot be considered really clean, in the sense that several phenomena take place: often, in case of interviews, some seconds of speech in the original language are played before the translation starts; commercials are often in English; there is the presence of music; sometimes a particular piece of news has not been translated yet in all of the languages, so that some channels may contain the original audio (in another language).

Every day we record one hour of audio stream: we first

[1] http://www.euronews.com

[2] https://www.eu-bridge.eu/

language	recording period	#videos	speech duration	#ref words	#rec words	#common words	aligned speech second (first) run
Arabic	01/01/13 - 31/05/13	4406	107:22:58	650,146	756,100	379,000	49:54:03 (42:39:46)
English	01/01/13 - 31/05/13	4512	112:18:29	973,210	1,032,727	699,850	63:02:34 (52:44:08)
French	01/01/13 - 31/05/13	4434	108:56:37	954,242	1,123,709	796,997	62:13:46 (57:55:07)
German	01/01/13 - 31/05/13	4438	108:33:23	809,289	896,387	653,372	61:46:27 (53:04:06)
Italian	01/01/13 - 31/05/13	4464	110:35:51	900,291	1,012,521	765,559	61:31:36 (54:01:00)
Polish	01/01/12 - 31/03/13	2626	58:32:03	350,729	454,977	278,854	27:42:35 (24:55:03)
Portuguese	01/01/13 - 31/05/13	4431	108:03:27	841,148	966,586	699,681	59:29:38 (43:24:16)
Russian	01/01/13 - 31/05/13	4418	107:42:24	714,363	828,060	611,347	60:49:15 (58:08:39)
Spanish	01/01/13 - 31/05/13	4465	109:16:50	939,408	1,053,255	797,698	63:29:23 (59:31:38)
Turkish	01/01/13 - 31/05/13	4387	106:30:31	683,041	764,329	556,760	60:52:55 (55:00:43)

Table 1: *Amount of Web data collected for AM training for the 10 EU-BRIDGE languages. The last two columns report respectively the number of common words resulting from the alignment among reference text and ASR output, and the resulting duration of speech (hh:mm:ss) after the second (first) run.*

extract the list of audio channels and then, for each language, its audio track which is stored at 16 kHz sampling rate. Given the particular structure of the stream, it is quite easy to segment the different audio tracks into news, because the boundaries among news are characterized by background noise or music over all of the channels. So, from a practical point of view, an effective and simple approach is to detect non-speech segments on each separate channel, and consider as probable news boundaries the non-speech intervals that are common to all channels. In this paper, TV data will be used only as part of the test set.

2.2. Data from the Web

Every day, Euronews stores in its portal videos and text describing the main news of the day. These latter are some of the news that are broadcasted via TV. The text associated with a news is sometimes a summary, sometimes a complete transcription, sometimes a partial transcription of the content of the news. By exploiting some information, it is possible to link a video to the same news in several languages. Approximately, Euronews publish on its portal about 30 videos per day for most of the languages. The average duration of a video is about 90 seconds, so in one month we are able to collect about 900 videos for a total duration of about 22 hours per language. However one of the EU-BRIDGE languages, Polish, is under-represented: it is currently not broadcasted via TV and the number of videos produced is less than 6 per day. This means that Polish data are different from the other languages.

To crawl news content, in the past we developed a program, called HLT WebManager [12], (free download from http://wit3.fbk.eu/ from the section "Tools"), that is able to download audio/video data, web pages, and to extract relevant text from them. It also provides a mechanism to find the parallel documents for each site.

3. Selection of multilingual data for AM training

In order to prepare material for AM training, it is necessary to collect a set of audio recordings together with their orthographic transcription. Our target was to collect about 100 hours of raw speech - including silence, music, etc. - for each language.

In order to obtain a reliable transcription of each news in an automatic way, we did the following [6]: perform ASR recognition on a news, align the ASR output with the text associated with the news, and retain only the speech segments where there is a perfect match. The amount of retained data mainly depends on two factors: the effectiveness of the ASR for the given language and the reliability of the transcription, and in our case ranges from about 45% to about 60% of the material.

We repeated the whole process twice; in both cases the LMs were trained only on the texts downloaded with the videos. The first run was performed using AMs trained at FBK in the past, exploiting composite corpora and performing quite differently. In some cases (Portuguese and Polish) we used multilingual AMs, trained on speech coming from 6 other languages, which are part of a language identification system [13]. The aligned speech resulting from the first run was then used to train a new AM for each language, which was finally used to perform the second run. As expected, in all cases the amount of reliable speech increased.

Details about the AM training data for the 10 languages are reported in Table 1, while a more accurate description of the selection process can be found in [1].

4. Multilingual data for ASR evaluation

To evaluate ASR performance it was decided to prepare and manually transcribe about 4 hours of speech for each language. Ongoing transcription is currently carried out by various EU-BRIDGE partners and, after some unexpected delay, is now scheduled to be completed by the end of June, 2014. Transcription guidelines consider usual phenomena like false starts, hesitations, speaker change, etc. Particular care was posed to the identification of segments of foreign speech (e.g. Arabic speech in the Italian data), quite common in Euronews data.

We collected more than 2 hours from the Web and nearly 2 hours from the TV, see Table 2 for details. Concerning the Web data, we acquired videos from the first week of July 2013, then a subset of these was selected, according to the following criteria:

- for each language, 120 files were selected, roughly corresponding to 155 minutes of audio, that could correspond to about 2 hours of true speech.
- precedence was given to news common to all languages, which resulted to be true for all languages except Polish. As a result, the same news compose the test sets for 9 languages.

From the TV we recorded data from the beginning of Octo-

language	recording	Web	speech	recording	TV	speech	total
	period	videos	duration	period	videos	duration	duration
Arabic	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
English	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
French	01/07/13 - 07/07/13	120	02:34:29	01/10/13 - 09/10/13	99	01:41:33	04:16:02
German	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Italian	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Polish	01/04/13 - 09/08/13	212	03:03:37	01/10/13 - 31/10/13	93(*)	01:28:50	04:32:27
Portuguese	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Russian	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Spanish	01/07/13 - 07/07/13	120	02:34:29	01/10/13 - 09/10/13	99	01:41:33	04:16:02
Turkish	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53

Table 2: *Amount of Web and TV data collected for development and evaluation for the 10 EU-BRIDGE languages. As Polish is currently not broadcasted via TV, Polish (*) comes from YouTube Euronews channel. Last column reports the whole duration of the speech data (hh:mm:ss).*

ber, 2013, keeping about 15 minutes per day. These data were automatically segmented into news, then news boundaries were manually checked and possibly corrected. For Polish , currently not broadcasted, we had to download more videos from the Web.

5. ASR baselines in 10 languages

For this experiments, we had to prepare baselines for all the languages. In particular, for each language, we had to:

- refine text cleaning and normalization procedures;
- prepare basic text processing (number processing, acronyms normalization, etc.);
- implement a rule-based phonetic transcriber.

Part of these scripts were implemented at FBK in the last few years (for instance [5, 14]), others were implemented ad hoc. We have to thank researchers from the Polish-Japanese Institute of Information Technology for their support for Polish text processing and for providing phone transcriptions for the Polish lexicon.

AMs were trained on the selected part of the training data resulting after the first run (see Table 1, very last column) while trigram LMs were trained on the text data collected along with the videos: their size is reported in Table 1, column labeled *#ref words*. Details about these procedures can be found in [4].

The transcription system used in all the experiments is based on several processing stages, briefly described here:

Segmentation, classification and clustering. The speech signal is divided in segments, based on a voice activity detector. The segments are mapped to several classes by a GMM classifier, and grouped in homogeneous clusters according to a BIC criterion. The clusters are used by the following acoustic normalization procedures. **Acoustic features extraction.** From the waveform, a sequence of 52-dimensional feature vectors is extracted, including 13 mel-scaled cepstral coefficients and their first, second and third derivatives. **Unsupervised acoustic features normalization.** The feature vectors undergo a first stage of normalization, computing a specific CMLLR transform for each segment cluster, with respect to a 1024-Gaussians GMM trained on the whole training set. **HLDA projection.** The 52-dimensional normalized feature vectors are projected in a 39-dimensional space, by means of an HLDA transformation. **First decoding step.** A first decoding step is performed on the resulting acoustic features, applying an AM based on tied-states cross-word triphone HMMs and an n-gram LM. This hypothesized word sequence is used as a supervision for the following supervised normalization. **Supervised acoustic features normalization.** The feature vectors are processed to perform a further normalization based on CMLLR transforms, this time exploiting the approximate transcription output by the first decoding step and a set of tied-states cross-word triphone HMMs with a single Gaussian per state. **Second decoding step.** A second decoding is performed on the normalized features, providing the final output.

	PP	OOV	WA
Arabic Dev	5986	39.1%	62.8%
Arabic Eval	5176	34.7%	58.3%
English Dev	481	4.3%	74.7%
English Eval	509	4.2%	79.5%
French Dev	168	4.1%	76.7%
French Eval	175	4.6%	76.8%
German Dev	574	6.6%	76.4%
German Eval	517	6.0%	79.2%
Italian Dev	299	5.6%	81.9%
Italian Eval	250	4.8%	83.5%
Polish Dev	756	11.5%	71.7%
Polish Eval	676	8.5%	81.6%
Portuguese Dev	409	10.0%	64.2%
Portuguese Eval	420	6.8%	72.0%
Russian Dev	1185	11.7%	65.0%
Russian Eval	1113	10.4%	66.6%
Spanish Dev	315	6.1%	88.5%
Spanish Eval	309	6.2%	84.9%
Turkish Dev	1311	14.9%	64.8%
Turkish Eval	1683	15.6%	64.4%

Table 3: *Results for the 10 baselines of the EU-BRIDGE dry run on the Euronews data. Perplexity, OOV rate and Word Accuracy are reported.*

5.1. Results

As previously said, we collected about 2 hours from the Web and about 2 hours from the TV channel for each language. Manual transcriptions are still under way and in January 2014 we performed a dry run, restricted to EU-BRIDGE partners, on a portion of that data (about half an hour for development and

	Ara	Eng	Fre	Ger	Ita	Pol	Por	Rus	Spa	Tur	correct
Arabic	83		1		2						96.5%
English	4	92	2		1				1		92.0%
French	1		97		1					1	97.0%
German	1		1	97					1		97.0%
Italian	1				98				1		98.0%
Polish	3				1	92		3		1	92.0%
Portuguese					3		94		3		94.0%
Russian	1						1	96	1	1	96.0%
Spanish									100		100.0%
Turkish	1									99	99.0%

Table 4: *LID results with the multilingual ASR system: correct identification rate is 96.15%.*

about half an hour for evaluation, for every language). Results for the 10 baselines for dry run dev set and dry run eval set are reported in Table 3, along with perplexity (PP) and Out of Vocabulary words (OOV) computed with the small LMs described above. As the quality of the data among the various languages is comparable, ASR performance depends on two main factors: AMs and LMs. As for AMs, their goodness could be roughly associated with the percentage of speech material selected from the training data, which in turn depend on the initial AMs used to perform the first iteration. LMs, on the contrary, depend mainly on the language perplexity, higher for inflected languages.

6. Language Identification task

We performed spoken Language IDentification (LID) experiments on part of the Euronews corpus prepared for ASR. Target languages were Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish. For each of the 10 languages, about 2 hours of speech were selected for training data while 986 speech segments, about 100 segments per language, were selected for evaluation. The 2 hours training data came from the Euronews training data, while test data came with manual annotation and transcriptions and were part of the dev set of the dry run evaluation; they are composed of manually selected segments having duration from 5 to 15 seconds.

The LID system is in fact a multilingual speech recognizer which exploits a multilingual set of triphone Hidden Markov Models (HMMs) and a multi-language N-gram LM to decode speech in an unknown language [13]. Multilingual triphone HMMs are based on a set of phones where the same phone (e.g. /a/) is shared among different languages. To this end, a common phonetic alphabet is used to produce phonetic transcriptions of words for each language. Texts amounting to about 2M words for every language are used to train a multi-language 3-gram LM with a fixed 5K lexicon for each language. At this stage, each word comes with its phonetic transcription in its language and is preceded by a LID label (e.g. deu:nicht, eng:home, fra:attaque, ita:cittadinanza). An ASR is then built capable to output words with attached a LID label. The multilingual speech recognizer generates the recognition hypothesis performing two decoding passes interleaved by AM adaptation. Finally, a majority filter is applied on the recognition hypothesis to assign a unique LID label to a whole speech segment.

The LID error rate obtained with the proposed system on the defined test set was 3.85%. In Table 4 we report the confusion array for the 10 languages.

7. Data availability

In 2013 the EU-BRIDGE consortium signed and agreement with Euronews, which gives to EU-BRIDGE partners the right to use Euronews material for research purposes, and to exchange it within the project. Concerning the availability of these data for the whole research community, in 2014 Euronews agreed to make it available for research purposes. At present, part of the data are available as AM training data for the ASR multilingual evaluation benchmark for IWSLT 2014[3]. An overview of the last IWSLT Evaluation Campaigns can be found in [15, 16].

8. Conclusions

In this paper we describe the design and collection of a multilingual speech corpus and report the first recognition experiments on it, in particular ASR and Language IDentification on a subset of manually transcribed data.

Speech and text data in 10 languages (Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish) come from the portal Euronews and were acquired both from the Web and from TV. For each language, the corpus is composed of about 100 hours of speech for training (60 for Polish) and about 4 hours, manually transcribed, for testing. We applied light supervision to the training data, obtaining a subset of the whole corpus with reliable transcriptions, that were used to train AMs. Small LMs were built exploiting the reference text coming along with the video. Linguistic processing was implemented to hande numbers and to provide phonetic transcriptions. 10 baselines were prepared - one for each language - and performance were evaluated on a subset of the test data. A LID system was implemented and evaluated, capable to recognize words belonging to different languages in a continuous stream. Part of the corpus is available for research purposes within the multilingual ASR benchmark for IWSLT 2014.

9. Acknowledgements

This work has been partially funded by the European project EU-BRIDGE, under the contract FP7-287658. The author wish to thank the EU-BRIDGE partners who contributed to transcriptions; Danijel Korzinek and Krzysztof Marasek from the Polish-Japanese Institute of Information Technology for their support for Polish text processing; Christian Girardi from FBK for its valuable help in downloading and cleaning text and video from the Web. A special thank to Walid Chamak from Euronews who gave us permission to use the corpus.

[3] https://sites.google.com/site/iwsltevaluation2014

10. References

[1] R. Gretter, "Euronews: a multilingual speech corpus for asr," in *Proceedings of LREC*, Reykjavik, Iceland, May 2014.

[2] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005, pp. 79–86.

[3] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *INTERSPEECH*, 2002.

[4] D. Falavigna and R. Gretter, "Cheap bootstrap of multilingual hidden markov models," in *Proceedings of INTERSPEECH*, Firenze, Italy, August 2011.

[5] A. Bisazza and R. Gretter, "Building a turkish asr system with minimal resources," in *Proceedings of First Workshop on Language Resources and Technologies for Turkic Languages*, Istanbul, Turkey, May 2012.

[6] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16(1), pp. 115–129, 2002.

[7] T. Schultz and A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme sets," in *Eurospeech*, Rhodes, Greece, 1997, pp. 371–374.

[8] T. S. D. Koll and A. Waibel, "Japanese LVCSR On the Spontaneous Scheduling Task with JANUS-3," in *Eurospeech*, Rhodes, Greece, 1997, pp. 367–370.

[9] J. Kohler, "Multilingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," in *ICSLP*, Philadelphia, 1996, pp. 2195–2198.

[10] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C. Lee, "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR," in *ICASSP*, Taipei, 2009, pp. 4333–4336.

[11] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[12] C. Girardi, "The hlt web manager," FBK, Trento, Italy, Tech. Rep. 23969, 2011.

[13] D. Giuliani and R. Gretter, "Esperimenti di identificazione della lingua parlata in ambito giornalistico," in *AISV*, Venezia, Italy, January 2013.

[14] A. Bisazza and R. Gretter, "Building an arabic news transcription system with web-crawled resources," in *Proceedings of the 6th Language and Technology Conference (LTC 2013)*, Poznan, Poland, December 2013.

[15] Federico, M. and Cettolo, M. and Bentivogli, L. and Paul, M. and Stüker, S., "Overview of the iwslt 2012 evaluation campaign," in *Proceedings of IWSLT*, Hong Kong, December 2012.

[16] Cettolo, M. and Niehues, J. and Stker, S. and Bentivogli, L. amd Federico, M., "Report on the 10th iwslt evaluation campaign," in *Proceedings of IWSLT 2013*, Heidelberg, December 2013.