Jacek ŚLIMOK, Jan KOTAS
SILESIAN UNIVERSITY OF TECHNOLOGY, FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE
Akademicka Street 16, 44-100 Gliwice, Poland

Evaluation of speech corpora for speech and speaker recognition systems

B.Sc. Jacek ŚLIMOK

Student of Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, in the field of Electronics and Telecommunications. Interested in various aspects of security problems, as well as ways of improving it.

e-mail: jacek.slimok@gmail.com

B.Sc. Jan KOTAS

Student of Automatic Control, Electronics, and Computer Science on the Silesian University of Technology in the field of Electronics and Telecommunication. Interested in microcontrollers, programming, and signal processing.

e-mail: kotas.janek@gmail.com

Abstract

Creating advanced speech processing and speech recognition techniques involves the need of working with real voice samples. Access to various speech corpora is extremely helpful in such a situation. Having this type of resources available during the development process, it is possible to detect errors quicker, as well as estimate algorithm parameters better. Selecting a proper voice sample set is a key element in the development of a speech processing application. Each speech corpus has been adapted to support different aspects of speech processing. The goal of this paper is to present available speech corpora. Each of them is shown in the form of a table. The tables contain the description of features helpful in choosing a suitable set of voice samples.

Keywords: speech recognition, speech processing, speech corpora.

Wykorzystanie baz mowy do testowania systemów rozpoznawania mowy oraz mówcy

Streszczenie

Tworzenie zaawansowanych technik przetwarzania oraz rozpoznawania mowy wiąże się z koniecznością pracy z rzeczywistymi próbkami głosu. Dostęp do różnorodnych zbiorów sygnałów mowy jest w tej sytuacji niezwykle pomocny. Posiadając tego typu zasoby, możliwe jest szybsze wykrywanie błędów, jak również lepsze oszacowanie parametrów algorytmów. Celem niniejszego artukułu jest zaprezentowanie dostępnych zbiorów próbek głosu. Dostępne bazy mowy różnią się między sobą między innym jakością, warunkami nagrywania oraz możliwymi zastosowaniami. Część baz zawiera rejestrowane rozmowy telefoniczne, z kolei inne zawierają wypowiedzi zarejestrowane przy użyciu wielu mikrofonów wysokiej jakości. Wykorzystywanie publicznych baz danych ma jeszcze jedną ważną zaletę - umożliwia porównywanie algorytmów stworzonych przez różne ośrodki badawcze, wykorzystujące tę samą metodologię. Uzyskiwane wyniki są prezentowane w postaci benchmarków, co umożliwia szybkie porównywanie opracowanych rozwiązań. Z tego powodu, wybór odpowiedniej bazy mowy jest kluczowy z punktu widzenia skuteczności działania systemu. Każdy ze zbiorów został przedstawiony w formie tabeli. Tabele zawierają opis cech pomocnych podczas wyboru odpowiedniego zbioru próbek głosu.

Słowa kluczowe: rozpoznawanie mowy, przetwarzanie mowy, bazy mowy.

1. Introduction

Creating advanced speech processing and recognition systems requires dozens of real voice samples. Instead of creating numerous variations of different accents, researchers may opt for using speech corpora, obtainable from multiple sources. Additionally, each possible voice sample set has been collected for specialized applications, focusing on different aspects of speech or speaker recognition. The available test results of other systems based on each corpus enable better evaluation of a system being in development. This paper describes some of the most popular speech corpora.

2. Corpora overview

This section describes speech corpora, created by numerous organizations. These collections of speech samples are gathered from different sources, such as telephone conversations, broadcast conversations or even broadcast news and consist of both male and female speakers in a wide age range. The mix of the attributes mentioned above as well as multiple languages allow creating a wide variety of possible combinations of voice samples. There is a wide variety of available speech corpora, most of which are not available for free, however they are often offered to members without a fee during specific year of membership. Most of the current speech corpora can be reached through The Linguistic Data consortium.

The Linguistic Data Consortium (LDC) is an open consortium of universities, corporations and research laboratories, formed in 1992. At first its role was to be a repository and a distributor of language resources. Currently the LDC creates, as well as distributes, a wide variety of language resources. It also supports research programs and language-based technology evaluations.

2.1. NIST 2000 Speaker Recognition Evaluation Corpus

NIST Speaker Recognition Evaluation is a part of yearly evaluations performed by NIST. They are designed to be as universal as possible, so they are useful to almost any research with independent speech recognition.

The 2000 evaluation focuses on four tasks:

- One speaker detection – this task focuses on checking if one, specific speaker is talking during speech segment
- Two speakers detection – this task is similar to the first one, except it focuses on both sides of conversation (eg. Telephone call).
- Speaker tracking – this task focuses on finding the intervals during which the speaker is talking during a conversation
- Speaker segmentation – this task focuses on finding the intervals when unknown speakers are talking. The number of speakers may not be known.

The yearly NIST Speaker Recognition Evaluation Corpus is aimed at multiple aspects of speech processing, all of which have been mentioned in the table above. Multiple systems are being tested using this corpus every year and the results of the best systems are published. An example of such a test is a two-speaker detection task using same sex segments, for which the best system has reached an error rate of 13.1% (year 2000, landline) [2].

Tab. 1. NIST 2000 Speaker Recognition Evaluation Corpus features
Tab. 1. Cechy bazy mowy NIST 2000

Languages	English
Source	telephone
Format	SPHERE, 8-bit mlaw
Length	148.9 hours
Member Fee	$0 (for 2001 members)
Non-Member Fee	$2 400
Recordings	10,328
Speakers	-
Application	speaker identification, speaker segmentation and tracking, speaker verification, speech recognition

2.2. TIMIT Acoustic-Phonetic Continuous Speech Corpus

The goal of the TIMIT was to provide data for the acquisition of acoustic and phonetic information in order to evaluate and improve automatic speech recognition systems. TIMIT was created by Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI), and was sponsored by Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO). It contains 6300 sentences spoken by 630 speakers (438 male, 192 female). All 8 major dialects regions of the United States are covered. The TIMIT also includes hand verified transcripts for all sentences [3].

Tab. 2. TIMIT Acoustic-Phonetic Continuous Speech Corpus features
Tab. 2. Cechy bazy mowy ciągłej TIMIT

Languages	English
Source	microphone
Format	1-ch PCM, 16kHz
Length	-
Member Fee	$0 (for 1993 members)
Non-Member Fee	$250
Recordings	6300
Speakers	630
Application	speech recognition

The TIMIT Speech Corpus has been adapted to support development of phoneme recognition systems. The recognition error rate of 24.6% has been reached by using a recurrent neural network, although it is described as one not being significantly different from other methods [4]. The recognition accuracy using this corpus has improved by about 13% over last 20 years. In 1990 the accuracy has reached around 66% and then went up to 75% in 1994. Since then, it has improved slowly and is currently estimated to be about 79% [5].

2.3. YOHO Speaker Verification Corpus

The YOHO database was created in 1989 under a US Government contract, and was not public until 1994. It contains many high-quality recordings, which can be used in speaker verification for security purposes. YOHO contains 1.5 Gigabytes of data that includes combination lock phrases (e.g. 12-45-32). It was collected over three months in four enrollment sessions per subject. There are 138 speakers (108 male and 30 female). Recordings were saved in PCM, with 8000Hz sample rate.

Tab. 3. YOHO Speaker Verification Corpus features
Tab. 3. Cechy bazy rozpoznawania mówcy YOHO

Languages	English
Source	telephone
Format	1-ch PCM, 8kHz
Length	-
Member Fee	$0 (for 1994 and 1998 members)
Non-Member Fee	$1 000
Recordings	1932 sessions, 24 phrases per session
Speakers	138
Application	speaker verification

The main purpose of the YOHO was to check whether a particular speaker verification system performed at 0.01% False-Accept and 0.1% False-Reject, with 75% confidence, understood as a likelihood that a given verification result matched the input. It is also checked, if a given speaker verification system passed the test with at least 50% probability. The YOHO can only be used in fixed-phrase speaker authentication research [6].

2.4. Mixer 6 Speech Corpus

Mixer 6 Speech was developed by Linguistic Data Consortium (LDC). Recordings contain 4410 public telephone calls, and multiple microphone sessions (1425). All 594 speakers have English as their native language. Each one of them was asked to complete 15 calls. Multiple microphone sessions used 14 microphones. In this part each person was asked to perform the following activities:

- Repeating questions (less than 1 minute)
- Informal conversation (about 15 minutes)
- Transcript reading (about 15 minutes)
- Telephone call (usually 10 minutes)

Telephone recordings were saved in 8kHz 2 channel SPHERE files. Microphone recordings are presented as 16kHz 1-channel FLAC/WAV files [7].

Tab. 4. Mixer 6 Speech Corpus features
Tab. 4. Cechy bazy mowy Mixer 6

Languages	English
Source	telephone, microphone
Format	1-ch PCM, 16kHz
Length	15,863 hours
Member Fee	$0 (for 2013 members)
Non-Member Fee	Not Available (Members Only)
Recordings	4 410
Speakers	594
Application	speech recognition

The Mixer 6 Speech Corpus has been created with speech recognition in mind. An example of such use can be HASR (Human Assisted Speaker Recognition) in NIST SRE10. The task was to test - given two speech fragments - whether they were spoken by the same person, which was a common verification trial. For HASR1 trials (across 20 systems) the cumulative miss rate (results of all tested systems combined together) reached about 38% and false rate at 47%. HASR2 (across 8 systems) obtained the miss rate of 35% and false acceptance rate of 41% [8].

2.5. Greybeard Speech Corpus

Greybeard has been developed by LDC and is a set of recorded telephone conversations, gathered in October and November 2008 and its aim is to gather voice samples from people participating in other speech collections in the past. It consists of 4680 calls, which make up approximately 590 or recordings, all in English. Each recording - after uploading - has been split into a 2-channel SPHERE-format file for each conversation and then manually verified for presence of background noise, cross-channel echo, as well as other difficulties and overall audio quality. This evaluation documentation is available upon obtaining the corpora. The main purpose of this speech sample set is to support speaker identifying system development.

Tab. 5. Greybeard Speech Corpus features
Tab. 5. Cechy bazy mowy Greybeard

Languages	English
Source	telephone
Format	2-ch PCM, 8kHz
Length	590 Hours
Member Fee	$0 (for 2013 members)
Non-Member Fee	$7 500
Recordings	4680
Speakers	172
Application	speaker identification

The main goal of the Greybeard Speech Corpus is to collect voice samples from past participants in order to gather longitudinal data about them. Such nature of the corpus enables research and development of long-term speaker recognition systems that takes the voice changes to ageing into account [9].

3. Evaluation methodology

During the development process of various speech processing and speech recognition systems, it is important to be able to assess the resources required to achieve the most effective results. Because of this, different speech corpora are created for different applications, which in turn reflects their content. All the mentioned speech sample sets have been evaluated by presenting the features that are most helpful for a person faced with a problem of selecting the right solution. Recording attributes that have been shown include: source, format, length, number of recordings, number of speakers and used languages. Aside from those, a fee has also been included, which should be useful for projects with limited budget. The information included in the parentheses under the member fee row indicates the year during which members were eligible for a given fee. Finally, an application presumed by its creators has been listed.

4. Summary

The paper has presented various speech corpora. The above rundown constitutes only a small subset of all available corpora and is in no way a complete collection. Only speech corpora obtained by the authors' University have been described. The shown comparison results can be useful in choosing a suitable set of speech samples during development of speech processing or speech recognition techniques.

This work was supported by The National Centre for Research and Development (www.ncbir.gov.pl) under Grant number POIG.01.03.01-24-107/12 (Innovative speaker recognition methodology for communications network safety).

5. References

[1] Rabiner L. R., Schafer R.W.: Introduction to Digital Speech Pro-cessing, Foundations and Trends in Signal Processing, Vol. 1, Nos. 1–2 (2007) pp. 1–194.
[2] Przybocki M., Martin A.: NIST Speaker Recognition Evaluation Chronicles, Speaker and Language Recognition Workshop, 2006.
[3] Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N.: DARPA TIMIT Accoustic-Phonetic Continuous Speech Corpus, February 1993
[4] Fernandez S., Graves A.: Schmidhuber J.: Phoneme recognition in TIMIT with BLSTM-CTC, April 2008.
[5] Lopes C., Perdigao F.: Speech Technologies, Chapter 14: Phoneme Recognition on the TIMIT Database, June 2011.
[6] Campbell, J.P., Jr.: Testing with the YOHO CD-ROM voice verification corpus, May 1995.
[7] Brandschain L., Graff D., Cieri C., Walker K., Caruso C., Neely A.: The Mixer 6 Corpus: Resources for Cross-Channel and Text Independent Speaker Recognition.
[8] Greenberg C., Martin A., Brandschain L., Campbell J., Doddington G., Godfrey J.: Human Assisted Speaker Recognition (HASR) in NIST SRE 10, July 2010.
[9] Kelly F., Drygajlo A., Harte N.: Speaker Verification with Long-Term Ageing Data, Proceedings 2012 5th IAPR International Conference on Biometrics (ICB), March-April 2012.

otrzymano / received: 24.03.2014
przyjęto do druku / accepted: 02.05.2014 *artykuł recenzowany / revised paper*