ELSEVIER

Speech Communication 31 (2000) 265–270

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

POLYCOST: A telephone-speech database for speaker recognition ☆

J. Hennebert [a,*,1], H. Melin [b], D. Petrovska [a], D. Genoud [c]

[a] *CIRC, EPFL, 1015 Lausanne, Switzerland*
[b] *KTH, TMH, SE-100 44 Stockholm, Sweden*
[c] *IDIAP, 1920 Martigny, Switzerland*

Received 19 August 1998; received in revised form 3 September 1999

Abstract

This article presents an overview of the POLYCOST database dedicated to speaker recognition applications over the telephone network. The main characteristics of this database are: medium mixed speech corpus size (>100 speakers), English spoken by foreigners, mainly digits with some free speech, collected through international telephone lines, and minimum of nine sessions for 85% of the speakers. © 2000 Published by Elsevier Science B.V. All rights reserved.

Résumé

Cet article présente une description de la base de données POLYCOST qui est dédiée aux applications de reconnaissance du locuteur à travers les lignes téléphoniques. Les caractéristiques de la base de données sont: corpus moyen à contenu varié (>100 locuteurs), anglais parlé par des étrangers, chiffres lus et parole libre, enregistrement à travers des lignes de téléphone internationales, minimum de neuf sessions d'enregistrement pour 85% des locuteurs. © 2000 Published by Elsevier Science B.V. All rights reserved.

Keywords: Speaker recognition; Databases

1. Introduction

We present in this paper an overview of the POLYCOST database dedicated to speaker recognition applications and assessments (Bimbot and Chollet, 1995). This database has been recorded within the framework of the European COST 250 action (Drafts, 1995) entitled "Speaker Recognition in Telephony" which has started in 1995 and ended in 1999.[2]

Speaker recognition research to date has focused primarily on wide-band speech. However, many applications, such as information services, merchandise ordering or financial transaction

☆ Expanded version of a talk presented at the Avignon RLA2C conference.

* Corresponding author. Present address: UbiCall Communications Inc., 1095 Market Street, Suite 3001, San Francisco, CA 94103, USA.

E-mail address: jean.hennebert@ubicall.com (J. Hennebert).

[1] This work was supported by the Office Fédéral de l' Education et de la Science (OFES) grant C95.0008, Switzerland, in the framework of the European COST 250 project.

[2] For more information, visit COST 250: Speaker Recognition in Telephony. http://www.fub.it/cost250/.

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00082-5

Table 1
Calls by countries and genders

Country	Male		Female		Total calls
	#	# Calls	#	# Calls	
Belgium-BE	5	35	3	24	59
Switzerland-CH	12	122	5	47	169
Denmark-DK	5	52	5	45	97
Spain-ES	5	51	5	51	102
France-FR	11	105	11	100	205
Ireland-IR	5	51	5	52	103
Italy-IT	5	49	5	50	99
Lituany-LI	1	9	0	0	9
Netherlands-NL	6	59	5	48	107
Portugal-PT	3	26	2	16	42
Sweden-SE	6	59	4	41	100
Turkey-TR	5	43	5	51	94
United Kingdom-UK	5	48	5	51	99
Total	74	709	60	576	1285

require the use of telephone speech. In Europe now, some speaker recognition dedicated databases are planned to be collected. The main project in this area is the SpeechDat project. [3]

The partners of COST 250 agreed to record a database dedicated to speaker recognition because

1. In Europe, there is a lack of speaker recognition databases over the telephone network.
2. A database recorded over the telephone line with a majority of non-native English speakers is potentially interesting to investigate and does not exist yet on the market.
3. A common database among the partners of the COST 250 action would permit easier assessments and comparisons of speaker recognition algorithms.

In Section 2, the contents and characteristics of POLYCOST are presented, focusing on the speaker set, the variabilities, the recording session frequency and the lexical contents of the database. Section 3 gives some details regarding the technical organisation underlining the recording procedure, the annotation work, the production and distribution of the database. Finally, a summary of the baseline speaker recognition experiments that are defined on POLYCOST is presented.

[3] Details on SpeechDat can be found on the project summary www page http://www2.echo.lu/langeng/projects/speechdat.

2. Characteristics and contents of the database

2.1. Speaker set

Speakers come from the European countries, members of the COST 250 action. Approximately 10 speakers per country were brought by each of the 13 partners. Table 1 gives the call breakdown of speakers per country and gender. No a priori control of the speaker distribution by gender, language and age has been done.

2.2. Variabilities

Besides the variabilities (spectral and temporal, intra and inter speaker) encountered in classical databases, other variabilities are present due to international telephone lines and the non-native language for most of the speakers. Table 2 gives the call distribution by languages and genders.

A survey was performed after the recording of the database in order to collect some statistics about age and phone sets variabilities. 103 answers were received from this survey. Table 3 gives the age distribution of the callers. It can be seen that age of subjects is concentrated in the region 25–35 years. Regarding the variability due to phone sets, about 80% of subject called from the same phone in all sessions. Table 4 shows statistics about the

Table 2
Languages represented and the number of speakers

Language	ISO-639	Male	Female
Italian	it	5	5
Portuguese	pt	3	2
Catalan	ca	2	1
Spanish	es	3	4
French	fr	23	16
Swedish	sv	6	4
Danish	da	5	5
Galician	gl	1	0
Dutch	nl	7	5
German	de	1	0
English	en	10	10
Turkish	tr	5	5
Lituanian	lt	1	0
Russian	ru	1	0
Arabic	ar	0	2
Macedonian	mk	0	1
Polish	pl	1	0
Total		74	60

Table 3
Age distribution of callers

Age	# Subjects
20–24	10%
25–29	34%
30–34	26%
35–39	15%
40–44	4%
45–49	6%
50–54	3%
55–59	2%

Table 4
Number of distinct phone sets used by the callers throughout their sessions

# Phones	# Subjects
1	79%
2	17%
3	3%
4	1%

number of distinct phone sets used by callers throughout their sessions.

2.3. Recording session frequency

Ten recording sessions were spread between February and April 1996, with a minimum spacing of three days between the sessions. Prompt sheets for every session were automatically sent to each speaker by fax or e-mail. Speakers were asked to make the call as soon as they received the prompt sheet in order to regulate somehow the intra-speaker variability. Table 5 gives the population of speakers by number of recording sessions. 85% of the speakers have recorded nine sessions or more.

Table 5
Population of speakers by number of recording sessions

# Sessions	# Speakers
1	1
3	2
5	1
6	3
7	9
8	4
9	14
10	75
11	20
12	3
14	2

2.4. Lexical content

One session is set up of 15 prompts (Hennebert et al., 1995) including 10 prompts with connected digits uttered in English, 2 prompts with sentences uttered in English and 2 prompts in mother tongue. One of the prompts in mother tongue is dedicated to free speech. The choice of the spoken material was driven by the interests of the different sites participating in the COST 250 project.

Individual client codes composed of seven digits were given to each speaker.

2.4.1. Session detail

English:

- 4 prompts spread during the session in which the speaker pronounces his 7 digits client code;
- 5 prompts distributed during the session in which the speaker pronounces a sequence of 10 digits (the same from session to session and from speaker to speaker);
- 2 prompts in which the speaker pronounces the sentences: "*Joe took father's green shoe bench out*" and "*He eats several light tacos*";

- 1 prompt in which the speaker is supposed to give his international phone number;

Mother tongue:

- 1 prompt in which the speaker gives his name, second name, gender (female/male), town and country;
- 1 prompt with free speech in which the speaker is asked to describe his environment or tell what he has done during the day. Speakers were kindly suggested not to say the same thing from call to call.

3. Technical organisation

3.1. Recording procedures

The database was collected through the European telephone network. The recording has been performed with ISDN cards on two XTL SUN platforms with an 8 kHz sampling rate. The acquisition platforms were located in Lausanne and Martigny, Switzerland.

Each speaker was asked to enter on the telephone pad its individual client code at the beginning of each session. This procedure allowed the automatic classification of the incoming calls by simple Dual Tone Multiplexed Frequency (DTMF) detection. Manual classification has been used in the case of no DTMF or wrong DTMF PIN code ($\sim$10% of the database). The client code sequences are 4–3 codes selected using a Reed Solomon coding which allows error correction and detection of false sequences.

3.2. Annotation of the database

Annotation is usually defined as the segmentation and labelling (or transcription) of speech databases. Experts agree on the general definition of these two processes: segmentation refers to the identification of stretches of speech signal; and labelling refers to the process of aligning a symbolic description of some speech items (labels) to the signal itself (den Os, 1997).

In the POLYCOST case, the English prompts are fully annotated in terms of word boundaries. The mother tongue prompts are just labelled at the word level with no segmentation. In both cases, the SpeechDat recommendations were used while performing the annotation. [4]

The annotation procedure of the English prompts was semi-automatic (Petrovska et al., 1996). The underlying idea was that, if an estimation of labels and stretches is available, the manual annotation work can be consequently reduced. The annotation procedure was then performed in two iterative steps. An initial annotation is first obtained by an automatic recognition procedure using word Hidden Markov Models (HMMs). Only utterances in which the recognised labels correspond to what speakers were supposed to utter are used for the second step. In the second step, the labelling is verified and the word stretches are manually adjusted while listening to the signal. The verified utterances are then used to re-train the HMMs that are becoming more accurate. Iterating these two steps improved the quality of the HMMs and increased the quantity of automatically annotated utterances. Finally, the remaining utterances that were not recognised were fully manually annotated.

3.3. CD-Rom organisation

The database is split into two CD-ROM discs. The speech files contain A-law data according to *ITU G.711* [5] standard (A-law coded samples, 8 kHz sampling rate, 8 bits/sample) with no file header.

3.4. Distribution of the database

A beta version of the database has been distributed with no charge to the members of the COST 250 action. This version had no annotations and some speech files were wrongly classified.

[4] More informations regarding the SpeechDat recommendations can be found under http://www.phonetik.uni-muenchen.de/SpeechDat.html.

[5] For more information on A-law encoding, visit International Telecommunication Union. http://www.itu.ch/.

POLYCOST will be available to users outside COST 250 through the European Language Resources Association (ELRA [6]) after the end of the COST 250 action.

4. Baseline experiments

In order to define a common ground for speaker recognition experiments on the POLYCOST database, a set of four baseline experiments has been defined (Melin and Lindberg, 1996). Results for one or more of those experiments should always be included when presenting evaluations made on the database. By including these results and by presenting the differences introduced in new experiments, a comparison between systems tested on different sites is made possible. This section presents a short summary of the ideas behind and the definition of the baseline experiments. Details on the experiments can be found in (Melin and Lindberg, 1996). For more information on speaker recognition system definition and assessment, the reader is referred to (Bimbot and Chollet, 1995).

Three of the experiments are speaker verification tasks and the fourth is a closed-set speaker identification task. The tasks are: (1) text-dependent speaker verification (SV) on a fixed password sentence, (2) text-prompted SV on digit sequences, (3) text-independent SV on free speech in the subject's mother tongue and (4) text-independent speaker identification on the same free speech. The experimental conditions in the four experiments were chosen to keep experiments realistic, well-defined and easy to implement. As far as possible, the four experiments have been defined with equal conditions. For instance, the same speakers and sessions are always used in the test phase.

In the baseline experiments, 110 speakers who have at least five available test sessions are used both as clients and simulated impostors. The total number of available test sessions for true-speaker tests is over 660 and there are around 12,000 independent impostor attempts in the verification tasks. A set of 22 other speakers, one male and one female from 11 different countries, have been set aside for use as an off-line database. The off-line database can for example be used to build world-models.

In general, annotation files provided with the database may not be used in the baseline experiments. The exception is at enrolment in experiment 2 (text-prompted digits), and with use of the off-line material.

5. Conclusion

Intra-, inter-speaker, language, and country variabilities are the keys for the understanding of speaker differences. POLYCOST is a database dedicated to speaker recognition applications over the telephone network. In its current development, the database presents the following characteristics: (1) more than 100 English speakers, mostly non-native, (2) utterances composed mainly of digits with some free speech, (3) recording through international telephone lines, and (4) more than eight sessions per speaker spread over a two-month period. A set of four baseline speaker recognition experiments has been defined on the database in order to facilitate cross-site comparisons of algorithms. The database will be available through ELRA at the end of the COST 250 action. Its cost should be low since it has been produced by a joint effort of the COST 250 members.

Acknowledgements

We are very grateful to all the COST 250 participants who have contributed to the recording of this database in giving us useful advice for the procedure and in bringing the voluntary speakers. As well, we would like to thank Telia InfoMedia Respons AB, Sweden, who sponsored the printing and distribution of the first version of POLYCOST to the countries participating to the COST action.

Warm thanks are going to Valérie Deillon, Paola Scarpini and Claudiu Budusan who gave

[6] More information under http://www.icp.inpg.fr/ELRA/fr/home.html.

their contribution to the deadly boring manual annotation work of the English prompts. Finally we thank all the well-meaning persons who gave their time to produce the labelling of the mother tongue prompts.

References

Bimbot, F., Chollet, G., 1995. EAGLES Handbook on Spoken Language Resources and Assessment, Chapter: Assessment of speaker verification systems. Mouton de Gruyter, Berlin.

den Os, E., 1997. EAGLES Handbook on Spoken Language Systems, Chapter: Spoken language system and corpus design. Mouton de Gruyter, Berlin.

Draft Minutes of the Inaugural Meeting, 1995. Cost 250: Speaker recognition in telephony, Technical Report, European COST Action, Rome, January.

Hennebert, J., Petrovska Delacrétaz, D., Melin, H., Genoud, D., Polycost v1.0 home page. http://circwww.epfl.ch/polycost.

Melin, H., Lindberg, J., 1996. Guidelines for experiments on the polycost database. In: Proceedings of a COST 250 Workshop on Application of Speaker Recognition Techniques in Telephony, Vigo, Spain, November, pp. 59–69.

Petrovska, D., Hennebert, J., Genoud, D., Chollet, G., 1996. Semi-automatic hmm-based annotation of the polycost database. In: Proceedings of a COST 250 Workshop on Application of Speaker Recognition Techniques in Telephony, Vigo, Spain, November, pp. 23–26.