TERASPEECH’ 2000 : A 10,000 SPEAKERS DATABASE

Marie-José Caraty, Claude Montacié

LIP6 - Université Pierre et Marie Curie
4, place Jussieu - 75252 Paris cedex 05 – France

Marie-Jose.Caraty@lip6.fr, Claude.Montacie@lip6.fr

ABSTRACT

TeraSpeech is a bilingual database (i.e., English and French) developed in partnership with a French museum, le Musée des Sciences et de l’Industrie in Paris. A demonstration of vocal signature is the support of this data collection. Aiming at the validation of a quality plan, a scenario of the demonstration has been designed, and various protocols have been developed. The quality plan is presented as well as the solutions we found for its validation (i.e., scenario and protocols). The statistics of TeraSpeech are given. Three trends are examined for the perspectives : the validation, the exploitation and the research. Over a single year of the vocal signature exhibition, TeraSpeech’ 2000 is a collection of more than 30,000 sentences recorded from more than 10,000 visitors. The exposition on acoustics of the museum is planned for ten years. TeraSpeech is expected to be a collection of more than 100,000 speakers recorded over the same sound acquisition channel.

1. INTRODUCTION

For oral linguistics, spoken language resources are the basis of observations which are of a major importance for the applications to come i) in automatic speech recognition, ii) in speaker recognition, iii) in concatenative synthesis, and iv) in language recognition. In the field of automatic speech recognition, besides the fact that in a statistical approach more and more spoken data are required for the improvement of the recognizers : an issue is still the study of the speech variability, and the ability to deal with within a recognizer. The definition we give here of the speech variability is the correspondence between various signals (i.e., speech realizations) and a same linguistic unit. Representative spoken databases are crucial -for the training of acoustic models, and as long as the variability isn’t controlled -for the development of adaptation techniques of the acoustic models. Multiple factors explain the speech variability and make difficult the design of a representative spoken database.

The biologic nature of speech is the origin of an intrinsic variability. There are three main kinds of intrinsic variability : i) the inter-speakers variability related to the individual physiology of the speakers, ii) the intra-speaker variability related to the psychological state of the speaker while pronouncing the “discourse”, and to the elocution mode (e.g., read speech, prepared speech, spontaneous speech, conversational speech), iii) the contextual variability related to the phonetic or linguistic neighborhood. Another kind of variability is extrinsic to speech, related to the material and environmental recording conditions including -the microphone, - the networks, -the environment (i.e., qualification of the noises). Guided by the applications to come, these spoken databases are crucial for realistic applications such as voice-driven services (e.g., over fixed/mobile networks, in car environment) [1] [2]. The recording conditions having a high influence on speech signals to be recognized, specific databases are needed. Let’ s call the recording conditions, the fixed choices for each of the previous factors of variability (i.e., intrinsic and extrinsic). The combinational explosion of the number of possible recording conditions makes impossible the collection of a universal database allowing the research on variability without bias. A spoken database has necessarily a purpose guiding the various protocols of collection whose the recording conditions.

Many other issues arise for the data collection : i) the quality plan of the database (e.g., quality of the recorded signals, quality of the various annotations, normalization), ii) the content of the database (i.e., specification of the items to record), iii) the speakers recruiting with the problem of quality of representation (i.e., distributions), iv) the raw cost of the data collection (e.g., remuneration for the participants, human or automatic supervision of the collection), v) the validation cost of the database (i.e., human cost or development cost), and at last but not the least important vi) the development (i.e., software and hardware) of the data collector.

TeraSpeech is designed for the research and the development related to inter-speakers variability. Speakers are recorded over two fixed channels of speech acquisition. An efficient way of recruiting the participants is an exhibition in a French museum devoted to the Sciences : the participants are the visitors who test a demonstration of vocal signature. For a complete automation of the data collection, a scenario guides the interactions between the visitor and the station of demonstration. The visitor is recorded with an automatic supervision of the speech signals. During the learning phase of a visitor, the amount of data per speaker is calibrated in duration, and the spoken data are authenticated.
2. DATA COLLECTION DESIGN

An exhibition in a very well-known French museum is the support of the TeraSpeech data collection. Through Orphée [3], the vocal key we designed, we have developed in partnership with the museum a demonstration on speaker recognition. Whatever is the objective of the databases (e.g., temporal derive, children voices, oral languages characterization, advanced terminals, voice-driven services over networks), many crucial issues arise for a data collection: -the recruitment of the participants and -the cost related to the organization of the data collection. The attraction of a museum dedicated to Sciences, its high frequentation, the curiosity of visitors for testing a vocal signature demonstration, the agreement of more than 90 percent of the visitors to let their recordings for research purposes, participate all together to the success of this collection of speakers and speech material. In this part, we present the quality plan we have defined for the data collection and the scenario of a speaker recognition session we have designed to validate in part the quality plan.

2.1. Quality Plan of the Data Collection

For our own research on speaker recognition, a first objective was originally to collect spoken data of as many as possible different speakers recorded over the same sound acquisition channel. A second objective was obviously to minimize the cost of the collection. To suppress the cost of human supervisor, the data collector had to be automatic. A solution we have found, opportune, is a vocal signature demonstration in a museum.

In the purpose of the future use of the database, the quality plan for the spoken data collection has been decided as follows:
1. Quality of the speech signals : requirement of a sound level of signals within a scale of magnitude. [Q1]
2. Quality of authentication of the spoken data : labeling in terms of speaker belonging. [Q2]
3. Calibration of the spoken data : -of the amount of data per speaker, -of the duration of the speaker recordings. [Q3]
4. Minimal annotation : requirement of the graphemic transcription of the spoken data. [Q4]
5. Quality of the correspondence between the spoken data and the graphemic transcription. [Q5]
6. Normalization of the database : requirement for maintenance (i.e., post-processing) and reusability. [Q6]
7. Minimization of the costs of the data collection. [Q7]

For the validation of this quality plan, the scenario of the demonstration has been adapted and protocols related to the collection have been defined.

2.2. Support of the Data Collection

The principle of the data collector is to use the speech technology : i) to popularize the speech research, ii) to bring participants, iii) to supervise automatically the quality control of both signals and annotations, iv) to improve this quality by explanation of the control feed back itself. TeraSpeech is recorded from two stations of demonstration. For each station, the demonstration scenario supervises the control of quality.

The first Station A is devoted to the learning phase:
1. Choice of the spoken language : English or French. [LP1]
2. Authorization : visitor agreement on the voice recording and its use for scientific purpose. If the visitor doesn’t agree, the spoken data are deleted in the evening. [LP2]
3. Microphone test : this test allows the adjustment of the voice level and the position of the visitor in relation with the microphone. [LP3]
4. Recording : the visitor name is recorded. [LP4]
5. Name acquisition: the visitor enters his pseudonym from a keyboard displayed onto the screen and a trackball. [LP5]
6. Beginning of the learning phase : the wholehearted participation of the visitor is requested concerning an ambient sound without burst of laughter or shouting. A text and its author are prompted onto the screen, the visitor begins to read aloud the sentence as soon as he is ready. The quality of the recording is supervised by the system. The visitor is informed when the sound level is too low or too high, and if the utterance can be taken into account for the computation of his model. In case of a well fitted sound level the first model is computed, the visitor is informed that two other reading stages are necessary to validate his model. [LP6]
7. Ending of the learning phase : new reading stages occur until the system has assessed the learning phase as successful (i.e., authentication of three recordings). [LP7]
8. Printing of a personalized card : the card is obtained through an opening in the cabin. The card of the vocal signature of the visitor contains his pitch, his pseudonym, and the representation of the signal of the visitor name previously recorded. [LP8]
9. End of the session : the visitor is asked to get to the next cabin to check his recognition by the system. [LP9]

The second Station B is devoted to the recognition test:
1. Choice of the spoken language : English or French pressing a button on the desk. [RT1]
2. Test session: a vocal message asks the visitor either for speaking spontaneously or for reading aloud (e.g., the text written on the station desk). [RT2]

3. Recognition decision: a vocal message is displayed followed by the recording of the name (R of N) of the nearest speaker of the visitors database. The visitor names have been previously recorded (LP4). The content of the vocal message is the following:

In English, “You sound familiar, aren’t you?” “R of N”
In French, “Vous me rappelez quelqu’un, ne seriez-vous pas?” “R of N” [RT3]

Per month, more than 1,000 visitors are recorded. The exposition is planned for ten years. A large-scale speakers database is expected (i.e., more than 100,000 speakers) in French and in English language.

3. DATA COLLECTION PROTOCOLS

In the purpose of the quality plan validation, protocols have been involved for the development of the vocal signature demonstration. We first describe i) the text corpora selection, then ii) the recording conditions. The statistics of the content of TeraSpeech’2000 are finally described.

3.1. Text Corpora Selection

Three main objectives have guided the selection of the textual corpora the visitors are required to read: i) free copyright texts in English and in French are chosen [P1], ii) for an homo-geneous amount of the speakers recordings, the sentences are calibrated in duration [P2], iii) aiming at a fluent utterance, the sentences are selected according to a cue of fluency [P3].

The text corpora are selected from the Gutenberg and the ABU databases respectively for English and French. These electronic libraries regroup writings, most of them being literary books. Dating from more than fifty years or seventy years, these writings are free copyright. [P1]

For an equal representation of the speakers (i.e., balanced recordings duration), the sentences are selected to provide, when reading aloud, an acoustic signal of about 10 seconds duration. The characters are related a priori to the phonemes and to the utterance duration. A first criterion of duration is used to select the sentences. Any sentence having between 160 and 240 characters is a candidate for the corpus. [P2]

For an easy reading, a cue of fluency has been defined. Each word is ranked according to its frequency of occurrence in the textual database (i.e., the most frequent word has the rank value 1). The cue of fluency of a sentence is computed as the average of the rank of its words. The sentences are ranked according to their cue of fluency. In the hands of the museum, a last filtering is the censure of any sentence which might displease the visitors. [P3]

For both languages, 10,000 sentences are finally taken out the original database from the cues of duration and fluency, and with the agreement of the censure. The statistics are the following: 9,834 different words in English, 19,264 different words in French, 13 seconds average duration per sentence, with 3 seconds of standard deviation. According to these criteria, we give here the first and the last sentence we have selected in English and in French. The text is given such as it is prompted aloud. (LP6, LP7). The set of selected sentences can be changed.

<table>
<thead>
<tr>
<th>EN</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of course father might have put some there he had got since, or that money might never have been his at all, but it seemed as if it would be, because it was on his land. [SOUND 00312_01.WAV]</td>
<td>“Laddie”, by Gene Stratton Porter</td>
</tr>
<tr>
<td>But stand thou by me at this hour, and according to thy will make straight my path, that thy glorious and dreadful name may be glorified in me thy servant, because thou art blessed for ever.</td>
<td>“Barlaam and Josaph”, by St. John of Damascus</td>
</tr>
<tr>
<td>Je sais bien que ce n’est pas la même chose à l’égard du monde ; mais au mien il n’y a aucune différence, puisque je sais que c’est par vous qu’il est mort, et que c’est à cause de moi. [SOUND 00312_02.WAV]</td>
<td>“La princesse de Clèves” (1678) de Madame de La Fayette</td>
</tr>
<tr>
<td>Les vases du fleuve ensevelissaient ces vengeance obscurcs, sauvages et légitimes, héroïsmes inconnus, attaques muettes, plus périlleuses que les batailles au grand jour et sans le retentissement de la gloire.</td>
<td>“Boule de suif” (1880) de Maupassant</td>
</tr>
</tbody>
</table>

Table 1: Examples over 10,000 sentences of the “most fluent” sentence ($\#1$) and the “less fluent” sentence ($\#10,000$) to be pronounced in English and in French.

3.2. Recording Conditions

Spoken data are continuous read speech. The acoustic signals are recorded from a microphone Beyer at 32 kHz sample rate, and 16 bit sample quantification. The Pinnacle sound acquisition card is characterized by a good signal/noise ratio (85 dB). The cabins $A$ and $B$ are delimited by window glass partitions with an opening on the exposition. From this opening are coming the background noise of the exposition, the noises of the bordering demonstrations (e.g., synthesis exhibition), and sometimes -an overlap of voices happening when few visitors are speaking inside the cabin during a learning or a test phase. Consequently, the noise environment isn’t predictable.

4. VALIDATION AND STATISTICS

Design and development of the data collection work towards its validation. We resume here the various factors contributing to the validation for three quality specifications. Statistics of the TeraSpeech’2000 are finally given in English and in French.
4.1. Validation of TeraSpeech’2000

[Q1] Quality of the speech signals : i) the sound level is controlled in the steps LP3, LP6, LP7, ii) the protocol P3 for the choice of “fluent” sentences to pronounce helps the processing for the validation of annotation quality Q5, iii) more subjectively, it’s noticeable at hearing that the visitors enjoy and apply themselves to read aloud writings of famous authors. [Q2] Authentication of spoken data : i) the precision of the recording date and the steps LP6, LP7 assess the learning phase and will enlarge the collection when, using speaker recognition technology, it is assessed that the same speaker has pronounced the three required sentences. [Q3] Calibrations of the spoken data : controlled thanks to the protocol of texts selection P2 and the assessment of the learning phase LP7.

4.2. Statistics of TeraSpeech’2000

At nearly the end of one year duration exhibition, 11,650 visitors of the museum have tested the vocal signature demonstration and 93 % of them have accepted to let the signal material for scientific purposes. The description of the statistics of the content of TeraSpeech’2000 [http://www-apa.lip6.fr/PAROLE/TeraSpeech/] is as follows :

<table>
<thead>
<tr>
<th>TeraSpeech 2000</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td># Hours of Recording</td>
<td>17</td>
<td>106</td>
</tr>
<tr>
<td># Speakers</td>
<td>1,568</td>
<td>10,082</td>
</tr>
<tr>
<td># Sentences</td>
<td>4,655</td>
<td>30,113</td>
</tr>
<tr>
<td>Sentence Average Duration</td>
<td>13.6 s (±3.4)</td>
<td>12.9 s (±3.2)</td>
</tr>
<tr>
<td># Sentences per Speaker</td>
<td>2.96</td>
<td>2.97</td>
</tr>
<tr>
<td># Words</td>
<td>131,370</td>
<td>987,997</td>
</tr>
</tbody>
</table>

Table 3: Statistics of TeraSpeech’2000.

5. PERSPECTIVES

Three trends are examined for the perspectives : i) the database validation, ii) the exploitation, iii) the research.

5.1. Validation

A first quality specification has to be validated : the graphemic transcription of the recordings. The text prompted for the learning session is a reference of this transcription. However, the quality of the transcription needs a supervision. The automatic supervision is indeed more advisable than a human supervision. Our studies on script warping [4] are useful for this purpose : the correspondence between the text corpus and the audio signal is estimated by the probability of a HMM-based alignment of the script to the audio signal. Multiple strategies may be conceived to validate or to reject an utterance. A second validation has to be carried out on the data collector B. Though the specification of the sound acquisition channels A and B are the same, the channels are different. It’s necessary to adapt more precisely the channel B for the validation of the authentication of the spoken data.

5.2. Exploitation

TeraSpeech is a well fitted database for evaluation plans (e.g., speaker recognition and speech recognition). The selection and the dissemination of learning and development sets is helped by , i) the high number of collected speakers, ii) the calibration of the spoken data, iii) the law cost of the data collection.

5.3. Research

Additional annotations will be available with post-processing to come, it’s essential to pay attention to the maintenance and consequently to the normalization of the descriptors of TeraSpeech. For the moment, we use [5] as description language XML like in MPEG7 standard.

TeraSpeech enables many studies on account of its huge number of speakers, and the variety of the speakers thanks to the family visiting of the museum (i.e., children and parents). Among these studies : i) the typology and the topology in speaker recognition, ii) the training of acoustic models in speech recognition, iii) the concatenative synthesis. For the last points based on the script warping technique, a useful study concerns the variants of pronunciation and their statistics.

Consequently to our research on speaker recognition, we are bent on studying specific typology and topology : we will soon appeal to twins and to imitators. An advantage of TeraSpeech is the sound acquisition channel which won’t change for ten years. At last, the extension of TeraSpeech to other languages (i.e., Spanish and Portuguese) is under consideration.

Acknowledgments : Our sincere thanks to our partners D. Boutboul, F. Vallas, A. Rabussier and E. Chollet of le Musée des Sciences et de l’Industrie, and to CDVI for its financial support.

6. REFERENCES