The SIWIS French Speech Synthesis Database – Design and recording of a high quality French database for speech synthesis

Pierre-Edouard Honnet[1], *Alexandros Lazaridis*[1],
Philip N. Garner[1] *and Junichi Yamagishi*[2]

Abstract

We describe the design and recording of a high quality French speech corpus, aimed at building TTS systems, investigate multiple styles, and emphasis. The data was recorded by a French voice talent, and contains about ten hours of speech, including emphasised words in many different contexts. The database contains more than ten hours of speech and is freely available.

Index Terms: database, French corpus, speech synthesis, emphasis

1 Introduction

Research on text-to-speech (TTS) synthesis requires high quality speech data. For some languages, a lot of free resources that are usable for TTS purposes can be found. Especially, English speech databases have been developed and made available in the last few decades.

For the French language, it is difficult to find well annotated high quality speech resources. Although some multi-speaker databases exist, and one can find relatively good quality free French audiobooks online[1], to the best knowledge of the author, there is no free French speech corpus allowing building high quality speech synthesis systems. The main limitation of audiobook speech is that it is not segmented and, despite some tools such as *ALISA* [Stan et al., 2016], requires a lot of work to be well segmented.

In the context of the *SIWIS* project, a multilingual database containing French (among Swiss languages) speech from multiple speakers was recorded [Goldman et al., 2016]. This corpus however contains limited amount of data per speaker and would rather be aimed at adapting existing systems than training them. An interesting feature of this database is the availability of specific emphasis patterns, produced by several speakers and in several languages.

Few such databases exist with sentences containing emphasised words. In English, one of the recent Blizzard Challenge datasets provided a high quality somewhat expressive speech corpus [Karaiskos et al., 2008]. The speaker, *Roger*, was asked to utter some of the sentences with a specific emphasis on one or two words. While this approach allows covering all the possible diphones in an emphasised word, the context of this emphasised word is limited to a specific scenario.

To enable studies of emphasis intonation in a more variable context, we designed a speech corpus which contains emphasised words in many different contexts.

This report describes the text preparation, and the recording procedure. The content of the database is presented to show the amount of data in each style of speech.

2 Text Material

The text selected for recordings consists of six parts:

- **parl**: consists of 4,500 isolated sentences from French parliament debates.
- **book**: consists of 3,500 isolated sentences from French novels.

[1] P.-E. Honnet, A. Lazaridis and P.N. Garner are with Idiap Research Institute, Martigny, Switzerland.
[2] J. Yamagishi is with the University of Edinburgh, United Kingdom.
[3] e.g. *Candide*, by Voltaire freely available on LibriVox: `https://librivox.org/candide-by-voltaire/`

- **siwis**: consists of 75 sentences from the *SIWIS* database.
- **sus**: consists of 100 semantically unpredictable sentences.
- **emph**: consists of 1,575 sentences taken from the 4 other sets.
- **chap**: consists of a full book chapter.

These 6 subsets serve different purposes. A detailed explanation on the text collection and preparation, as well as the purpose of each set is given below.

2.1 parl

With the primary goal being the construction of TTS systems, the text was taken from debates between June 2012 and July 2015 at the French parliament. The main advantages of using such data are its free and open access, and its contemporary aspect: the vocabulary used in the debates reflects the current language in the society. The text data was downloaded from the French national assembly website[2]. A first selection was done to get the best possible diphone coverage, using a greedy algorithm. To ensure easy readability and remove ambiguity of pronunciation, we only considered sentences that contained common words; in practice, we ensured that the sentences only contained words that were among the 10,000 most common in French. The word frequency for the French language was estimated from recent French news articles available online from *Le Monde*, *Liberation* and *TF1 news*. A selection based on the sentence length, removing the shortest and longest sentences, allowed reducing the sentence set. Many sentences with the same recurrent structure were manually removed, for instance sentences like “La parole est à Monsieur David Douillet.”

2.2 book

In order to have copyright free material, five French books were selected from two authors: *Zadig* and *Micromegas* from Voltaire, and *Voyage au centre de la Terre*, *L’Île mystérieuse* and *Vingt mille lieues sous les mers* from Jules Verne[3]. After removing very short and very long sentences, a first random sentence selection was made. A manual checking was done in order to remove the sentence containing names whose pronunciation could be ambiguous.

2.3 siwis

This set corresponds to the sentences from the *SIWIS* database for which the speakers were asked to emphasise specific words [Goldman et al., 2016]. Recording the same data allows us to have parallel data from one database to the other. The sentences come from the Europarl corpus, which was designed to have a parallel meaning across languages [Koehn, 2005].

2.4 sus

These sentences were partly taken from the *SIWIS* database, for which 20 such sentences were selected and recorded. The remaining 80 sentences were generated on a sentence generator website[4]. The generator randomly produces sentences respecting grammatical rules. The set of generated sentences was manually checked to remove sentences which made sense. This subset was recorded for testing, as semantically unpredictable sentences are good candidates to evaluate intelligibility for instance.

2.5 emph

These sentences are aimed at studies on emphasis. In a similar fashion to the *SIWIS* database, this allows collecting the same sentences in 2 different styles: one in a neutral style, and one in which the speaker is asked to emphasise a specific word more than the rest of the utterance. The 1575 sentences consist of 800 sentences from **parl**, 600 sentences from **book**, all the 75 sentences from **siwis**, and all the 100 sentences from **sus**.

[2] The URLs follow the syntax `http://www.assemblee-nationale.fr/14/cri/2012-2013/20130001.asp`, where 14 stands for the 14th term of office, cri for integral report, 2012-2013 for the current session year, 2013 for the current civil year, 0 for ordinary session, 001 indicates that it is the first session of the year.

[3] The books are freely available on `http://beq.ebooksgratuits.com`

[4] `http://romainvaleri.online.fr/`

2.6 chap

This part is a full chapter taken from the French book *Vingt mille lieues sous les mers* from Jules Verne (Chapter III of the book). It amounts to approximately 1800 words. The fact that the sentences are not isolated like in the other sets makes it attractive for studies on higher level prosody. The style of reading, because of the material which contains both dialogues and narration, makes this part of the database more expressive by design.

3 Speaker and Recording

The recordings were conducted by a professional voice recording agency called Voice Bunny[5]. The speaker was selected from a pool of native female French voice talents, by the first author, who is a native French speaker. The instructions were provided to the speaker as follows:

- For **parl**, **book**, **siwis**, **sus**: read each sentence in an isolated manner, with a long pause (> 2 s.) between each sentence.
- For **emph**: read the sentence with focus on the indicated word, e.g.:
  "**Lourde** [read out this bold word with emphasis] *erreur, madame la ministre !*"
- For **chap**: read the full chapter in an expressive manner, without long pauses between sentences.

The sentences with emphasis were recorded after their neutral version, in order not to influence the speaker to reproduce the patterns that they were asked to produce in that case. Finally, the book chapter was read in one session, in order to have the dependencies that one can expect when reading a long text, e.g. the gradual downdrift of intonation along a paragraph, and reset when starting a new paragraph.

The data was recorded using Studio Projects B1 microphone, and provided in 44.1kHz mono 16 bits. Adobe Audition[6] was used for processing. 30 minutes of recordings generally required 90 minutes of processing, editing and checking from the voice actress. In total, 23 sessions were necessary to complete the recordings.

4 Database content

Table 1: Amount of speech data recorded.

Style	# of sentences (sessions)	Time (inc. silences)	Time (no silence)
parl	4500 (6)	4h02	3h44
book	3500 (11)	5h00	4h45
siwis	75 (1)	3.8min	3.5min
sus	100 (1)	4.5min	4min
emph	1575 (5)	1h33	1h26
chap	—[7](1)	10.5min	10.5min
Total	9750 (23[8])	10h54	10h13

Table 1 gives the amount of recorded speech in terms of utterances and time. The times without silences were estimated based on the automatic alignment performed on the contextual labels, and correspond to the times between the start and end of speech (meaning that the mid-sentence pauses are counted as speech). These labels were created using eLite[9] [Roekhaut et al., 2014]. The first five parts of the database were segmented by finding long silences, and keeping short silences before and after actual speech; the full chapter was not segmented.

Figure 1 illustrates the position of the emphasised words, both in an absolute and relative manner. The relative position was simply obtained by dividing the position by the total number of words, including intermediate silences. Many emphasised words are located in the second half of the sentence, but their

[5] `https://voicebunny.com`

[6] `http://www.adobe.com/products/audition.html`

[7] This part was not segmented in sentences.

[8] sus and emph were recorded in the same session as one of the book sessions.

[9] A webservice, available at `http://cental.uclouvain.be/elitehts/v1/`.

Table 2: Number of syllables in emphasised words.

# syllables	1	2	3	4	5	6	Total
# words	803	568	246	67	10	1	1695

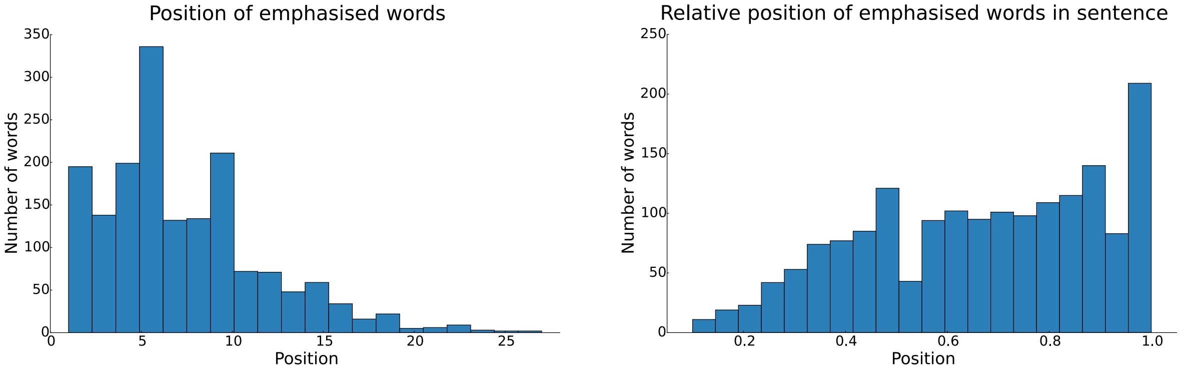

Figure 1: Emphasised word positions. Left: absolute position, right: relative position in the sentence.

absolute position is generally lower than 10, mostly because the majority of the sentences are short. There are in total 1695 annotated emphasised words, for 1575 sentences (some sentences had several emphasised words).

Table 2 gives the distribution of the emphasised words by number of syllables. 95% of the emphasised words contain 3 or fewer syllables.

When looking at the word-level context as defined by the HTS label format, the 1695 words correspond to 1450 different contexts (to have the same context, 2 words need to have exactly the same number of syllables, position in the phrase and utterance, part-of-speech, etc.).

5 Summary

This report presented the design, recording and content of a new freely available speech database. It is a high quality French database, containing speech from multiple styles. Its primary purpose is speech synthesis, but it also contains sentences with emphasis on specific words, in many contexts. The data is available with no restrictions on the usage, under Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

6 Acknowledgements

The authors would like to thank Blaise Potard for his help in the selection of phonetically balanced sentences. The database has been recorded with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

References

Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, and Junichi Yamagishi. The SIWIS database: a multilingual speech database with acted emphasis. In *Proceedings of Interspeech*, pages 1532–1535, San Francisco, CA, USA, September 2016.

Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo, editors. *The Blizzard Challenge 2008*, 2008.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Sophie Roekhaut, Sandrine Brognaux, Richard Beaufort, and Thierry Dutoit. eLite-HTS: A NLP tool for French HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 2136–2137, 2014.

Adriana Stan, Yoshitaka Mamiya, Junichi Yamagishi, Peter Bell, Oliver Watts, Robert Clark, and Simon King. ALISA: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech & Language*, 35:116–133, 2016.