A canadian french emotional speech dataset

Conference Paper · June 2018

DOI: 10.1145/3204949.3208121

Philippe Gournay
Université de Sherbrooke

A Canadian French Emotional Speech Dataset

Philippe Gournay, Olivier Lahaie, Roch Lefebvre
Speech and Audio Research Group
Université de Sherbrooke
Sherbrooke, Québec, J1K 2R1, Canada
Philippe.Gournay@USherbrooke.ca

ABSTRACT

Until recently, there was no emotional speech dataset available in Canadian French. This was a limiting factor for research activities not only in Canada, but also elsewhere. This paper introduces the newly released Canadian French Emotional (CaFE) speech dataset and gives details about its design and content. This dataset contains six different sentences, pronounced by six male and six female actors, in six basic emotions plus one neutral emotion. The six basic emotions are acted in two different intensities. The audio is digitally recorded at high-resolution (192 kHz sampling rate, 24 bits per sample). This new dataset is freely available under a Creative Commons license (CC BY-NC-SA 4.0).

CCS CONCEPTS

• Computing methodologies → Language resources

KEYWORDS

Emotional speech; Canadian French; speech dataset; digital recording

ACM Reference format:

P. Gournay, O. Lahaie, and R. Lefebvre. 2018. A Canadian French Emotional Speech Dataset. In *Proceedings of the ACM Multimedia Systems Conference, Amsterdam, Netherlands, June 2018 (MMSys'18)*, 4 pages. https://doi.org/10.1145/3204949.3208121

1 INTRODUCTION

An ever-increasing number of emotional speech datasets are available, with new datasets featuring various languages and content types being released virtually every year [1, 2]. These datasets are essential to affective computing [3] and more generally to the development of better human-computer interactions [4]. They are the starting point for fundamental research aimed at identifying which spectral and prosodic cues are involved in the expression of voice emotions [5], as well as for more applied research such as automatic recognition of speech emotions [6].

Despite all this, no emotional speech dataset had been recorded and made available in the Canadian French language until recently. This was a limiting factor for research in Québec as well as in the rest of the French-speaking world. In particular, this lack of available audio material made subjective experiments quite challenging, because it is generally preferable that the audio samples are in the native language of the human subjects participating in the experiments.

In this paper, we present the design and content of the first and newly released Canadian French Emotional (CaFE) speech dataset. The paper is organized as follows. The construction of the sentence set and the selection of the emotions are discussed in section 2 and 3, respectively. Some information about the actors is given in section 4. The recording setup is described in section 5 and the item selection and file naming is explained in section 6. Finally, conclusions are drawn and future work is discussed in section 7.

2 CHOICE OF SENTENCES

2.1 The sentence set

The number of sentences in the dataset had to be small enough so that the contribution of each actor could be recorded in a reasonable amount of time. At the same time, it had to be large enough to allow for a variety of research activities.

The sentences had to be emotionally neutral from a semantic point of view, yet well suited to be acted in various emotions. They also had to be reasonably easy to pronounce. Finally, they had to be composed of the same number of syllables so that neither one has more importance than the others.

From a phonemic point of view, the set of sentences had to be as complete and as balanced as possible. The French language is usually considered to be composed of 38 different phonemes. Two of these phonemes have been deliberately ignored: the first one because it is slowly falling into disuse (the long ɛ: as in the French words “fête” and “maître”) and the other one because of its particular nature (the ŋ as in parking, which comes from English and has very little representation in French [7]).

The resulting set of six different French sentences developed for the CaFE dataset, with their phonemic transcriptions and English translations, is given in Table 1. These sentences all have eight syllables.

Table 1: The CaFE set of six sentences with their phonemic transcriptions and English translations.

1	Un cheval fou dans mon jardin /ɛ̃ ʃəval fu dɑ̃ mɔ̃ ʒaʁdɛ̃/ One crazy horse in my garden
2	Deux ânes aigris au pelage brun /døz- ɑn ɛgʁi o pəlaʒ bʁœ̃/ Two embittered donkeys with a brown coat
3	Trois cygnes aveugles au bord du lac /tʁwa siɲ avœgl o bɔʁ dy lak/ Three blind swans by the lake
4	Quatre vieilles truies éléphantesques /katʁə vjɛl tʁɥi elefɑ̃tɛsk/ Four gigantic old sows
5	Cinq pumas fiers et passionnés /sɛ̃k pyma fjɛʁ e pasjɔne/ Five proud and passionate pumas
6	Six ours aimants domestiqués /siz- uʁs ɛmɑ̃ dɔmɛstike/ Six domesticated caring bears

2.2 Phonemic distribution

The CaFE set of sentences from Table 1 contains a total of 115 phonemes. The distribution of these phonemes is given in Table 2 with, for each phoneme, an example of a word featured in the dataset that contains it.

The distribution of phonemes in the French language has been the subject of many studies, each one conducted on a specific spoken and/or written corpus [8]. The Wioland distribution of French phonemes [9] is provided in Table 2 for a comparison between the French language in general and the CaFE set of sentences in particular.

The Wioland distribution was measured on 77,702 phonemes from a large corpus combining spoken (radio broadcast) and written (literacy texts) French. Considering the diminutive size of the CaFE set of sentences, it is clearly impossible that the two distributions match exactly. But, as can be seen from Fig. 1, the deviation between the CaFE and the Wioland distributions is typically in the order of plus or minus one phoneme, which is quite satisfactory.

Table 2: Number of phonemes in the CaFE set of sentences, with examples of words containing the phonemes, compared to the Wioland distribution [9].

Phoneme	Example	No.	CaFE	Wioland
Consonants				
b	bord	2	1.74%	1.08%
d	jardin	5	4.35%	4.24%
f	fou	3	2.61%	1.38%
g	aigris	2	1.74%	0.56%
k	quatre	5	4.35%	3.75%
l	cheval	6	5.22%	5.89%
m	pumas	4	3.48%	3.91%
n	ânes	2	1.74%	3.09%
ɲ	cygnes	1	0.87%	0.14%
p	pelage	3	2.61%	3.88%
ʁ	jardin	9	7.83%	7.58%
s	cygnes	7	6.09%	5.75%
ʃ	cheval	1	0.87%	0.61%
t	trois	5	4.35%	5.39%
v	cheval	3	2.61%	3.00%
z	deux-ânes	2	1.74%	1.55%
ʒ	jardin	2	1.74%	1.57%
Semi-vowels				
j	vieilles	3	2.61%	1.76%
w	trois	1	0.87%	1.03%
ɥ	truies	1	0.87%	0.37%
Vowels				
a	cheval	9	7.83%	8.11%
ɑ	ânes	1	0.87%	0.05%
e	passionnés	5	4.35%	5.28%
ɛ	aimants	6	5.22%	5.55%
ə	cheval	3	2.61%	3.39%
i	cygnes	5	4.35%	5.08%
œ	aveugles	1	0.87%	0.44%
ø	deux	1	0.87%	0.51%
o	au	2	1.74%	1.97%
ɔ	bord	3	2.61%	1.28%
u	fou	2	1.74%	2.62%
y	pumas	2	1.74%	2.01%
Nasals				
ɑ̃	dans	3	2.61%	3.21%
ɛ̃	jardin	3	2.61%	1.16%
œ̃	brun	1	0.87%	0.54%
ɔ̃	mon	1	0.87%	2.27%

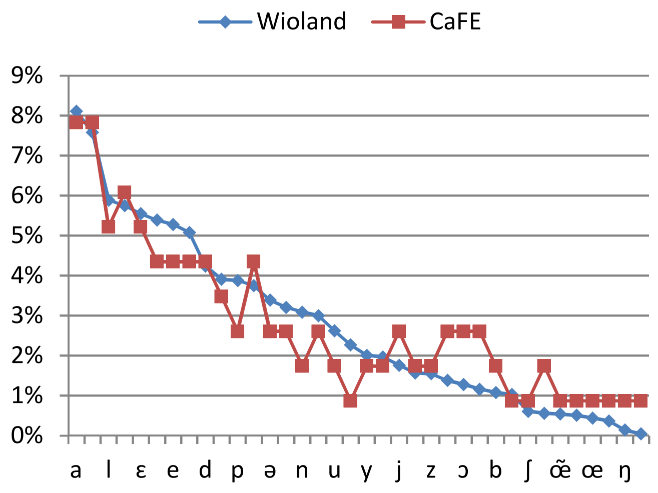

Figure 1: Phonemic distribution in the CaFE set of sentences compared to the Wioland distribution [9].

3 CHOICE OF BASIC EMOTIONS

Many different theories have been proposed for human emotions, with some continuous and some discrete [10]. Discrete theories assume various numbers and types of basic emotions [11]. We have chosen to follow Ekman's theory, which defines six basic emotions: sadness, happiness, anger, fear, disgust and surprise [12]. This adequately covers the spectrum of emotions that can be conveyed by speech. This is also very similar to what can be found in other emotional speech datasets such as [13]. Other theories suggest other basic emotions, such as love, hope, courage or interest, but these are much more difficult to share using oral speech only.

Some of these basic emotions are richer and harder to define than others. For example, surprise is often tinged with fear. During the recording sessions, the actors were instructed to produce emotions that were as independent as possible.

Each of the six basic emotions was acted in two different intensities (low and strong). For the interpretation of these two terms, the actors were guided by a couple of verbal indications. They were told for example that low-intensity sadness would feel like having a lump in the throat, while a strong-intensity one would be almost bursting into tears. Also, one neutral emotion was recorded (obviously, in only one intensity).

4 THE ACTORS

Table 3 gives the gender and age of each of the twelve actors who participated in the recordings. They were all Canadian French speakers with a Québec French accent. They were also all professional or amateur actors with a background in cinema, television or theater.

Table 3: Gender and age of each of the twelve actors who contributed to the CaFE dataset.

Actor	Gender	Age
01	M	46
02	F	64
03	M	18
04	F	50
05	M	22
06	F	34
07	M	15
08	F	25
09	M	42
10	F	20
11	M	35
12	F	37

5 RECORDING SETUP

The recording sessions took place one actor at a time, in a professional soundproof room. The sound was recorded by a Yeti Pro USB microphone from Blue Microphones. This microphone was mounted on a tripod equipped with a pop filter. The actors were standing in front of the microphone and allowed to move freely. The microphone was set in cardioid mode and connected via a USB cable to a remote ACER Swift 3 laptop. The recording was done at the highest possible resolution (192 kHz sampling rate, 24 bits per sample).

6 ITEM SELECTION

6.1 Item selection

The typical duration for a recording session was one hour per actor. Over one session, three to five takes were typically recorded for each sentence, in each emotion, and at both intensity levels. Out of these, the best take was selected according to criteria of quality of acting (particular attention was paid to incorrect pronunciation, hesitations and excessive mouth noise) and quality of recording (to minimize clothing noises and inadequate sound levels). The selected take was then segmented, leaving between three-quarters and one second of speech inactivity (background ambiance) before and after the speech utterance.

Although the original recording was done at the 192 kHz sampling rate and with a resolution of 24 bits per sample, a filtered and downsampled version at the more common format of 48 kHz sampling rate, 16 bits per sample has also been produced and is made available.

6.2 File naming

The CaFE speech dataset contains a total of 936 audio samples, each one in a separate audio file. The file naming convention is as follows:

AA-E-I-S.wav

where AA is a two-character field that gives the actor number (01 to 12), E is a one-character field that indicates the emotion according to Table 4, I is a one-character field that indicates the intensity level (1 for low, 2 for strong), and S is a one-character field that gives the sentence number (1 to 6). The intensity field is always 1 (low) for the neutral emotion (N).

Table 4: Identification letter for the emotions in the CaFE dataset.

Letter	French	English
C	Colère	Anger
D	Dégoût	Disgust
J	Joie	Happiness
N	Neutre	Neutral
P	Peur	Fear
S	Surprise	Surprise
T	Tristesse	Sadness

7 CONCLUSION

This paper presented the newly released Canadian French Emotional (CaFE) speech dataset. This dataset includes six different sentences, pronounced by twelve actors, in six basic emotions plus one neutral emotion. The basic emotions are acted in two different intensities. This represents a total of 936 different audio samples.

This dataset is freely available under a Creative Commons license (CC BY-NC-SA 4.0) and can be downloaded from the following URL:

http://www.gel.usherbrooke.ca/audio/cafe.htm

or from the CaFE Zenodo repository:

https://doi.org/10.5281/zenodo.1219621

Future work is aimed at enriching this dataset, first by recording more sentences and more actors to make this dataset even more attractive for deep learning (which is known to require large amounts of training data). Then, by diversifying the content of the dataset, for example by including a number of children voices.

ACKNOWLEDGMENT

The authors would like to thank Professor Mathieu Nuth (Université de Sherbrooke) for selecting the actors and participating in the recordings.

REFERENCES

[1] D. Ververidis and C. Kotropoulos. 2006. Emotional Speech Recognition: Resources, Features, and Methods. *Speech Communication*, Volume 48, Issue 9, pp. 1162-1181, September 2006.

[2] M. El Ayadi, M. S. Kamel, and F. Karray. 2011. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*, Volume 44, Issue 3, pp. 572-587, March 2011.

[3] R.W. Picard. 1997. *Affective Computing*. The MIT Press, USA, 306 pages.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. 2001. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, Volume 18, Issue 1, pp. 32-80, January 2001.

[5] K. Sreenivasa Rao and S.G. Koolagudi. 2013. *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer Briefs in Electrical and Computer Engineering – Speech Technology, 126 pages.

[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. 2011. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication*, Volume 53, Issues 9-10, pp. 1062–1087, November-December 2011.

[7] H. Walter. 1983. La nasale vélaire /ŋ/, un phonème du français ? *Langue française*, n°60, pp. 14-29, 1983. Numéro thématique : *Phonologie des usages du français*, sous la direction de Henriette Walter.

[8] J. De Kock. 1983. De la fréquence relative des phonèmes en français et de la relativité de ces fréquences. *ITL – International Journal of Applied Linguistics*, Volume 59, Issue 1, pp. 1-54, 1983.

[9] F. Wioland. 1972. Estimation de la fréquence des phonèmes en français parlé. *Travaux de l'Institut de Phonétique de l'Université de Strasbourg*, Volume 4, pp. 177-204, 1972.

[10] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor. 1987. Emotion Knowledge: Further Exploration of a Prototype Approach. *Journal of Personality and Social Psychology*, Volume 52, Issue 6, pp. 1061-1086, 1987.

[11] A. Ortony and T.J. Turner. 1990. What's basic about basic emotions? *Psychological Review*, Volume 97, Issue 3, pp. 315-331, July 1990.

[12] P. Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, Volume 6, Issue 3, pp. 196-200, 1992.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A Database of German Emotional Speech. In *Interspeech'2005*, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005.