BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis

David Guennec, Hassan Hajipoor, Gwénolé Lecorvé, Pascal Lintanf, Damien Lolive, Antoine Perquin, Gaëlle Vidal

BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis

David Guennec ∗ *, Hassan Hajipoor* ∗ *, Gwénolé Lecorvé* ∗ *,*
Pascal Lintanf †*, Damien Lolive, Antoine Perquin, Gaëlle Vidal*

Univ Rennes, CNRS, IRISA, France
† Skol Vreizh, France
damien.lolive@irisa.fr, antoine.perquin@irisa.fr, gaelle.vidal@irisa.fr

Abstract

Breton is a minority language spoken in the Brittany region of France. Public initiatives are being undertaken in order to preserve the Breton language. As an effort toward that goal, we created a large Breton speech corpus and related automatic annotation tools. The corpus contains 20 hours of reading aloud for both a male and a female Breton speaker. Then, end-to-end text-to-speech synthesis systems are built. Subjective evaluation suggests that the systems are able to reproduce the voices of the original speakers faithfully.

1. Introduction

Development of resources and technologies for under-resourced languages is an urgent task for endangered languages such as Breton. It is an Indo-European language, and the last continental Celtic language still in use. Just over two hundred thousand people use it today in everyday life whereas active speakers were more than a million in 1950, as a result of French acculturation, industrialisation and new mobility habits.

Celt culture was widespread over main parts of Europe during centuries in the ancient classical times, and only survived in British Isles in two main forms: Gaelic (spoken in Ireland, Scotland and the Isle of Man), and Brythonic, through Welsh, Cornish, and also Breton that extended from Great Britain to Brittany during the first millennium. Breton evolved into different dialects, but orthography standards were defined and are used by new learners. However, it appears to be difficult to follow a suitable pronunciation, including the traditional accentuation.

New speech technologies are a chance to preserve minority languages that are threatened to be lost due to the low amount of speakers. Works toward that preservation goal have already been conducted for other Celt languages, such as Text-to-Speech synthesis systems [1]. To the best of our knowledge, ours is the first work to attempt the same for Breton since early attempts using diphone concatenation [2].

A public initiative launched in 2018 by the Regional Council of Brittany aimed to fill a gap in current provision and availability of voice services in Breton. The PUBLIC OFFICE for the BRETON LANGUAGE (Breton: *Ofis Publik Ar Brezhoneg*) assigned technological aspects of this mission to us. The Breton-language publishing house SKOL VREIZH assisted us by taking part in linguistic work, text data supply and speakers management.

The goal of this work is two-fold. First, the construction of a Breton speech corpus and associated tools that could be used for speech related tasks. Second, the validation of that corpus by designing text-to-speech (TTS) synthesis systems that reproduce faithfully the voice of the recorded speakers.

The creation of the speech corpus was a three-step process. First, after linguistic specifications were defined to act as a reference point for a standard Breton pronunciation, the speech corpus creation started with the selection of texts to be used as prompts during the audio recording process. The resulting selection of texts is presented in Section 2. Then, text processing is needed. Section 3 introduces the protocol used for text normalisation, as well as the grapheme-to-phoneme (G2P) module we built to perform phonetisation. Objective evaluation of those systems shows good results overall, with errors occurring mostly on numerical expressions and liaison-between-words. Investigations suggest that, most of the time, errors are due to substitutions between labels referring to similar sounds. Finally, as described in Section 4, we were able to record a speech corpus which, after cleaning and processing, resulted in more than 20 hours of speech for each of two fluent Breton speakers (one male, one female). This corpus is longer than the 8 to 10 hours minimum needed to achieve good synthesis performance. It is also comparable to, or larger than what is commonly available for most low-resource languages [3]. We put emphasis on the variety of text sources during the corpus design to allow innovative applications in future works.

The creation of a large corpus of Breton speech allowed us to build TTS systems using state of the art neural architectures. Section 5 describes the structure and training of our TTS sytems. Namely, the conjunction of a Tacotron2 [4] and a ParallelWaveGAN model [5]. Finally, subjective evaluations presented in Section 6 show that our systems were highly rated by Breton speakers and were able to reproduce the voice of the original speakers faithfully. Interestingly, in the case of this Breton dataset, the speaker similarity perceived by listeners was found to be significantly better for grapheme inputs than for phoneme inputs.

2. Text Selection for Corpus Recording

In order to reach the first goal of creating a corpus that could be applied to different tasks on the Breton spoken language, we first needed to select texts to be read during later recording sessions. This section introduces the method used to select those texts, as well as the resulting selection.

∗ David Guennec is now employed by ViaDialog

∗ Hassan Hajipoor is now a PhD candidate at University of Massachusetts

∗ Gwénolé Lecorvé is now employed by Orange Innovation

The Breton language has multiple orthographies recognized as standards in Brittany. Under the advice of expert Breton speakers, we only considered texts using the *peurunvan* orthography, because it gathers particularities of the four main Breton dialects (spoken in Kerne, Leon, Tregor, and Gwened counties of Brittany), and is used nowadays in most of the schools were Breton is still taught.

Then available texts were collected from different sources, and an amount of required text data, for 20 hours of reading aloud, was estimated from the speech rate recorded during the speakers casting. A 130,000 words corpus was built in order to have a diversified database in terms of forms and registers. As shown in Fig.1 and detailed in Table 1, it consists in a selection of texts from three main distinct sources:

- Journalistic sentences (9,676 utterances), from the Public Office for the Breton Language. Their linguistic register is mainly administrative, they are extracted from newspaper articles, exhibition presentation, websites as Wikipedia Breton, and also technical documentation (dispensers, time stamps, ...).

  Three-quarter of them constitute half of the whole corpus, and are dedicated to neutral speech. They were selected from a first provision, after exclusion of sentences showing more than 20 words, and aiming to have equal distribution of words in the different sentences. The last quarter of them is reserved for their specific content (abbreviation or French words), or dedicated to dictation speech for experiment.
- Literary (55 short works), from Skol Vreizh editing house, providing one sixth of the whole corpus through 16 works with dialog, 10 prose works, 5 tales and 23 poems.
- Everyday-life sentences (3,217 utterances), from Mozilla Common Voice database [6], providing another sixth of the whole corpus. Most of them are declarative, but there are also a lost of questions, and exclamations.

Other text data were defined to complement the corpus:

- Lists of symbols (131 utterances showing spelled letters and phonemes, and some typographic characters), town names (556), forenames (763);
- Cooking recipes (11 works, written by the female speaker's mother);
- Sentences showing non-speech sounds (135 utterances), built by inserting tags into randomly selected Mozilla Common Voice sentences;
- Sentences showing rare diphones (884 utterances), built according to the type of diphone target. Rare diphones were deduced from the diphone distribution of the ordinary journalistic sentences sub-set. Amongst the 1,849 theoretically possible diphones out of the 43 phonemes Breton set, 740 were missing, and 86 only appeared once. To meet this need five sub-corpus were defined:
  - 175 missing diphones, as they appear in the phonetised reference lexicon, could be added through a simple selection of words, and also some dedicated sentences were written by Skol Vreizh.
  - 428 missing diphones that do not appear in the lexicon could be added either by building two-words utterances where they appear in the liaison, or by inventing words.
  - 55 rare diphones could appear at least a second time in other sentences written by Skol Vreizh.

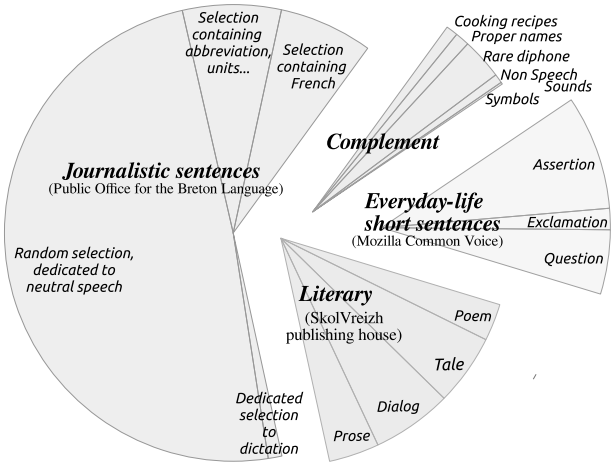

Figure 1: Text data global distribution

3. Text processing

Once the texts to be used during the recording sessions have been selected, further works were needed to be done on them. Namely, text normalization and phonetisation. This section introduces the implementation and evaluation of these systems.

As a pre-requisite for these two processes, we wrote or compiled, from different reference materials, normalisation and pronunciation rules, and also a phonetised word lexicon. Notably the collective dictionary *An Here*[7] and two books edited by Skol Vreizh, *Francis Favereau's dictionary*[8] in 2015, and grammars, as *Jean-Claude Le Ruyet*'s one[9] in 2012 were used.

A phonetised lexicon was built and phonetisation rules were implemented, both based on a selection of IPA symbols.

3.1. Text normalisation

Performance of automatic text normalisation relies on how much the system is adapted to data format (i.e. character encoding, punctuation), and how far it succeeds in producing full specific forms (abbreviations, units, morpheme, acronyms ...) and in some cases uses specific rules (number expressions, internet address or other). Part-Of-Speech information was extracted from dictionaries using Universal POS tags following hierarchy attributes, but it was not yet exploited.

A 1% Word Error Rate was automatically measured on a test-set of 1,495 utterances (9 words on average), randomly selected from a subset of long ordinary journalistic sentences. Errors impact 3.5% of sentences. 47% of word errors are wrong automatic translation of number expressions into their full written form. Three quarter of them are substitution when ordering words, the rest is wrong choice for numerical words, as they can depend on gender, and can be of different types (undefined or not, ordinal, ...). A few errors also occur with numbers in mutation situation (liaison, gender distinctive and disambiguation mutations change to close grapheme the first consonant of a word). The second cause of word errors (41%) is misinterpretation of the suspension point (written "pik" *("dot")* in the way of internet address spelling). Others word errors are wrong nor-

Journalistic
Example: 10 embregerezh eus ar vro vigoudenn o sinañ evit ar brezhoneg. (Ten companies from the Bigouden country are signing for the Breton language.)
- Random selection of sentences (49%), with balanced distribution of length $\leq$ 20 words,
- Selection containing specific forms (7%): abbreviation, units, mail, ...
- Selection with French words (6%)
- Selection to dictation speech (1%)

Literary
Example: Fresk e oa en damdeñvalijenn. (It was cold in the darkness.)
- Dialog dominant literary (6%)
- Tales (5%)
- Prose dominant literary (4%)
- Poems (2%)

Everyday life
Example: "Hag amzer hoc'h eus da lenn ar gazetenn diouzh ar mintin ?" ("Do you have time to read the newspaper in the morning ?")
- Declarative (8%)
- Question (4%)
- Exclamatory (2%)

Complement
Example: Ha pa zistrofe ar gouloù, <cough>e vefe re ziwezhat. (And even if the light came back, <cough> it would be too late.)
- Rare diphones (2.9%)
- Proper names (1%)
- Cooking recipes (0.8%)
- Non-speech sounds (0.6%)
- Linguistic symbols (0.1%)

Table 1: Text corpus distribution and examples

malisation of spelled letters (7%), and a few wrong processing of some words and abbreviations.

As a result, half of word normalisation error can easily be solved. The rest of the score depends on improvement of numeral expressions processing. It is indeed a particular challenge in the Breton language, with complex ordering, implying mutation rules and gender dependent article generation, and the need for a POS processing.

3.2. Grapheme-to-Phoneme

The new standard Breton pronunciation convention could cover specificity of three of the four main Breton dialects, but the one spoken in Gwened country was found too different to be taken into account. A set of phonetic labels was defined as shown in Table 2. It uses 45 IPA phonetic symbols including suprasegmental markers for primary stress and length (accentuation), and a tag for pause event. Once the text is normalised, phonetisation is performed by lexicon for words that are recognised, and by rules when the word is unknown, and also to deal with liaison-between words. Rules directly act on graphemes in context to predict the phonetic string. As no large phonetically annotated text is available, the only way to build a first G2P module was to define rules.

3.2.1. Stress marking evaluation

On a set of 5k isolated words an evaluation shows 97% of accuracy for stress labelling. For the moment, no work has been done on stress prediction at the sentence level. It should be conducted in the future.

Consonants													
nasal	m				n			ɲ	ŋ				
plosive	p,b				t,d				k,g				
sibilant fricative				s,z	ʃ,ʒ								
non-sibilant fricative		f,v								x		h	ʁ
palatal approximant							j						
lateral approximant					l		ʎ						
Co-articulated						w					ɥ		
Vowels													
close					i			y		u			
close-mid				e			o		ø				
open-mid				ɛ			ɔ		œ				
open	a	ɑ											
nasalised		ɑ̃		ɛ̃	ɛ̃	ĩ	õ		ỹ	œ̃		ũ	
Pause #	Primary stress '				Length mark ː								

Table 2: Standard Breton phonetic labels

3.2.2. Phonetic labelling evaluation on isolated words

A 2.2% Phoneme Error Rate was measured on a subset of isolated words, randomly selected from the Lexicon, but without using the phonetic information during the process. 273 rules were called, and it produced 461,816 phonetic labels.
Most of rules with high score of errors are general (lack of precision of the system). Example of general rule: *⟨ o ⟩ in other cases*

ensoc'hañ (word)
ensɔhɑ̃ (lexicon phonemes)
ensohɑ̃ (rules phonemes)

But PER is also impacted by some rules with bad contextualisation, notably for graphemes ⟨ e ⟩, ⟨ i ⟩, and ⟨ o ⟩. They suggest further investigation including validation of the lexicon. Example of contextual rule: *⟨ e ⟩ between ⟨ 3 letters or more ⟩ and ⟨ a consonant at the end of a word ⟩*

dour-bev (word)
durbew (lexicon phonemes)
durbɛw (rules phonemes)

3.2.3. Phonetic labelling evaluation on long sentences

A 7.36% Phoneme Error Rate was measured on a test-set of 50 long utterances (the mean number of word by sentence is 20). 4,104 labels were generated. Errors are mainly substitutions (218), then insertions (70) and deletions (14). In fact, most of errors are substitution between labels corresponding to similar sounds.

Substitution within pairs of consonants voiceless/voiced (shown in the appropriate zone in Table 2 with comma separator) are 22.8% of total errors. They notably concern /t/ /d/, /s/ /z/ and /k/ /g/. The system tends to generate labels for unvoiced sounds where they should be voiced. Most of them are final consonant wrong labelling in situation of liaison between words, where pronunciation shows *Sandhi* phenomena that slightly distances it from the writing. Here is an example of sentence with five sandhis, most of them correspond to a voiced phoneme :

Aet eo kuit evit ur veaj hir, noz ha deiz war ar mor.
'ɛː**d** ɛw 'kli**d** evi**d** œr 'veaʃ 'hiːr # 'noː**s** a 'dɛjz war ar 'moːr
(He left for a long journey, night and day on the sea.)

Substitution between close-mid vowels[1] and their open-mid corresponding are 23.8% of total errors (mainly /e/ /ɛ/). The rest of substitution are also for close sounds: /h/ /x/, /ɑ/ /a/, and nasalised/non-nasalised vocalic (respectively 7.3% 6.3% and 3.0% of total error).

Another noticeable type of errors is label insertion. Here again in a liaison situation, insertion of /t/ and /d/ are both 13.2% of total errors. They can happen in case of the elision (i.e. no pronunciation) of the final consonant. It often happen in Breton language for very common words in expressions, e.g. ⟨ mon**t** a-gleiz ⟩ ⟨digan**t** an⟩ ⟨oant ke**t** sikouret zo be**t** lakaet⟩.

Table 3 suggests a comparison, for pairs of close labels, between the sum of their distribution and the sum of their PER score. It confirms concentration of errors, notably on the pair /t/ /d/ and /e/ /ɛ/.

Groups of labels for close sounds	% total labels	% total errors
a ɑ	16.0	7.0
e ɛ	14.8	19.9
t d	9.6	24.5
s z	6.3	9.6
x h	1.9	7.9
Nasalised vowels	1.1	2.5

Table 3: Main phonetic label errors on a 50 sentences test-set

The phonetic system is less performing in processing sentences than it does for words, but most of errors concern substitution between labels for similar sounds.

Automatic phonetic strings were produced for the whole text corpus, to guide the speakers during the recording process. Despite prediction errors, it was found to be helpful in that context.

4. Speaker and recording framework

4.1. Casting

Two speakers were chosen after the casting of seven native speakers of Breton language. Several selection criteria were used, such as: speech intelligibility, dynamism, voice quality, and of course their flexibility to adjust their pronunciation to new conventions defined upstream as part of the project.

- Male: Pascal, 48. Native from Tregor county, he has a traditional background in terms of language. He was also linguistic expert within the project, and directly adapted his dialect to the standard. Pascal's speech is very stable in terms of intonation, accentuation and rhythm. In constraining situations his phonation tends to be fast, with long pauses.
- Female: Annaig, 66. Native from Tregor county, she has a more standard background. She had been for decades a well-known anchorwoman on TV in Breton Language. Annaig's voice is particularly good and lively sounding, and her speech shows very accurate articulation. Because of her background her efforts had to be addressed to accentuation emphasizing.

4.2. Recording

Recording took place in IRISA recording studio, using the following devices: acoustic cabin environment, DPA-4060 headset microphone, Lynx Studio Technology PCI/ISA AES 16 sound card with Aurora 8 A/N N/A converter. Signal was stored in the format 44.1 kHz 16 bits mono with Audacity software.

Speakers had to follow the exact recording script, including phonetic content. However, during the course of the project it was found more comfortable to leave some degree of latitude concerning mid-sentence pauses position (see below). In that sense, in addition to a very few files that are not common to both speakers, the two speech corpus are not exactly parallel. Indeed differences appears in the way speakers realize punctuation.

For each speaker, ten days of work were needed to record the whole twenty hours corpus. Half-day recording session were divided in three sets, that could produce 15 to 30 mn of final speech each. This task required efforts and preparation to adapt to new pronunciations, and for many utterances two or more shots were needed before reaching the target. Everything was recorded, including talks and commentaries made in-between takes.

4.3. Audio post-processing

A multi-step post-process was applied to produce final audio data. As the recording work was quite a difficult task for both speakers and the recording staff, audio split and indexation were postponed until after the recording sessions were done. Moreover, specific audio-work was done to reconstitute some utterances or moderate their tempo.

Once sets were recorded they were cleaned by selecting best versions among recording attempts and exclude commentaries. It sometimes lead to use different shots for the same utterance, when there was hesitations or errors that lead to repetitions.

Then loudness was normalized to -23 LUFS (Loudness Unit Full Scale) using measurement and recommendations given by the open-source tool *FreeLCS* [2], and globally applied to each cleaned recording set.

Finally, the fully prepared audio data sets were split into utterances using a python script that only extracts audio over a given RMS treshold (Root Mean Square measures amplitude in terms of its equivalent power content), and does not during long silences. Then audio utterances could be indexed with text data.

4.4. Issues with recording process and remediation

During the first recordings there was an issue with the speech rate of the male speaker when reading aloud the whole journalistic sub-corpus dedicated to neutral speech). He had received instructions to respect the punctuation in the text, but faced the difficulty of reading aloud a text that was dedicated to a written form and thus presented a deficiency of comma. As a result he spoke faster, as a strategy to catch his breath the soonest possible at the end of utterances. This phenomena also slightly impacted Pascal's voice features, as they are influenced by timing constraint.

Other factors could increase the speech rate of the male speaker for this part of the corpus that was recorded first: he was the first aloud reader and had to make efforts to follow the new pronunciation rules, and efficiency was a big concern. This could lead him to produce longer pauses to have time to think, and instinctively reduce speech length, to re-balance global time.

Concerning this data, as there was no time to repeat the work for the whole subset, it was found necessary to change the speed before further processing. On 41% of the male's corpus, speech

[1] vowel height depends on position of the tongue, lowered or raised

[2] http://freelcs.sourceforge.net/

speed was slowed down by increasing file length by 10%, without changing the pitch, in order to stay as close as possible to the original vocal timbre and tone (micro silence insertion by "change tempo" Audacity built-in effect standard algorithm). Fortunately the remaining part of the male speaker's recordings was done later. After having discussed on priorities and after lifting the constraints on punctuation, it was read in a stable and natural way, at a moderate speed. Annaig's speech was also recorded later, and didn't suffer from speed adjustment: she could read punctuation in a free way, and she could have prepared her speech before by listening to Pascal's version.

Final audio provides a 20h 35mn speech corpus for the male speaker, and a 21h speech corpus for the female speaker.

In the final database, normalisation was entirely conformed to speech, including punctuation. Concerning the phonetic string, a complete conformation to speech of 1% of the whole corpus was manually produced in parallel between each speaker. It's made up of linguistic symbols and proper names, plus 250 journalistic sentences. It specifically deals with recurrent or salient automatic errors.

5. Speech Synthesis for the Breton Language and for Speaker Reproduction

The second goal of our collaboration with the Regional Council of Brittany was to use the speech dataset described previously to develop a text-to-speech synthesis system. Our main concerns were not only to achieve high quality synthetic speech, but also to make sure that the synthetic speech reproduced faithfully the voice of our recorded speakers, and sounded like spoken by a native Breton speaker.

While works on multi-speaker TTS models are currently ongoing and show interesting promises, the speaker similarity between natural and synthetic speech is usually lower with multi-speaker models than single-speaker ones. Furthermore, they usually require to be trained on large amounts of speakers. In light of these observations, we opted to train independent single-speaker models for each of our recorded speakers.

Finally, in order for synthetic speech to sound like native Breton, the architecture of the TTS model needed to be language-independent. We decided to use the Tacotron architecture which has been shown to work on a wide number of languages.

5.1. Model Architecture

The architecture of the whole TTS system is comprised of a Tacotron2 model followed by a ParallelWaveGAN vocoder, as drawn on Figure 2.

Tacotron2 [4] is a model that predicts mel-spectrogram from a text represented as a sequence of either phonemes or graphemes. It is composed of an encoder that extract high level features from text. An attention model attends over those high level features to align them with the synthetic mel-spectrogram. Finally an auto-regressive decoder predicts the mel-spectrogram frame-by-frame using the concatenation of the output of the attention model and the last predicted frame. More precisely, the decoder uses 3 internal sub-networks. The first one, the Prenet, performs a non-linear transformation of the last predicted frame. The second one is a recurrent neural network (RNN) that predicts a mel-spectrogram frame from the concatenation of the outputs of the Postnet and attention model. Finally, the Postnet adds fine-grain details to the predicted mel-spectrogram frame through a residual connection.

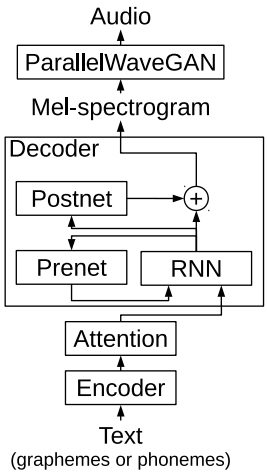

Figure 2: Architecture of the TTS system

The Tacotron2 model was created and trained using the ESPNET Toolkit [10]. The encoder is composed of a stack of 3 convolutional layer with 5 filters of 512 channels followed by a BiLSTM layer of dimension 512. The attention model is location-based, implemented using a convolution layer with 15 filters of 32 channels. The attention output are projected in a 128 dimensional space using a fully-connected layer. In the decoder, the Prenet, is a stack of 2 fully-connected layers of dimension 256 followed with ReLU activation functions. The RNN is a stack of 2 LSTM layers of dimension 1024. Finally, the Postnet is a stack of 5 convolution layer using 5 filters with 512 channels.

To convert the mel-spectrograms back to audio, a ParallelWaveGAN vocoder [5] per speaker was trained using the code and default hyper-parameters defined in Tomoki Hayashi's popular implementation [3].

5.2. Subset of the Corpus Used for Speech Synthesis

Experiments on TTS were conducted in parallel of the audio and text processing of the corpus. Thus, at the time of training, the data available to train our models were only a subset containing 60% of the corpus described in the previous sections.

Figure 3 describes the distribution of sentences according to the number of words they contain, with respect to the style of the textual content :

- Journalistic sentences: 7,925 sentences, 60,747 words, from the ordinary subset dedicated to neutral speech. The mean for number of words by sentence is 7.5. The mean length of audio samples is 4.19 and 4.65 for the male and female voice respectively.
- Everyday life sentences: 3,208 utterances, 17,458 words. Those sentences are shorter (mean 5.5 word/sentence), and show explicitly direct style linguistic content even for most of declarative forms (common expressions in everyday speech, direct style punctuation markers, use of personal pronouns and verbal forms). The mean length of audio samples is 2.83 and 3.10 for the male and female voice respectively.

At the time of the TTS experiments, part of the everyday life sentences were not yet processed for the female voice, lead-

[3] https://github.com/kan-bayashi/ParallelWaveGAN

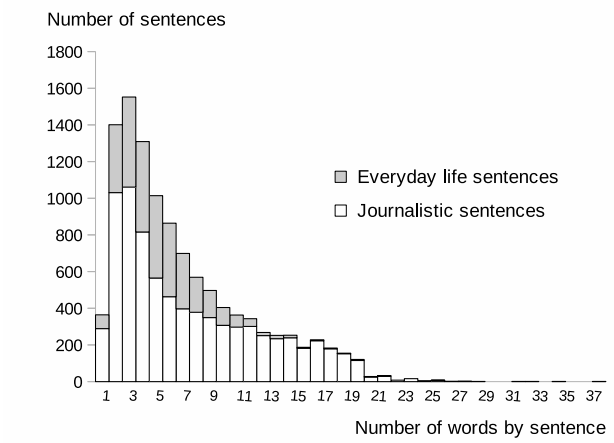

Figure 3: Words by sentence distribution in the 11,133 sentences finalised data subset

ing to a small mismatch in the amount of data available between the male and female voice.

For each textual prompt, the output of the text normalization process is used to train the grapheme-based model, while the output of the G2P process is used to train the phoneme-based model.

The audio pre-processing for the training of our models consisted in trimming the leading and trailing silences longer than 30ms and down-sampling the audio to 22050 kHz. Afterward, 80 dimensional mel-spectrogram were computed using a Hann window of size 1024 and a shift of 256.

5.3. Model Training

The audio samples described in section 5.2 were used to train and evaluate our speech synthesis systems. The dataset is divided between utterances spoken by the male and female speakers. It is further divided into subsets. Out of all samples available, 250 were kept for the test set and 250 were kept for the development set. This resulted in a total of 8862 and 10633 samples, respectively, in the training sets of the female and male voices.

We trained the Tacotron2 architecture on both graphemic and phonemic input independently, for both the male and female speakers. The models were trained for 200 epochs using Adam as an optimizer with a learning rate of 1.10^{-3} and batches of size 32. We used an early stopping mechanism with a patience of 20 epochs, leading the training to stop early around the 50 epochs mark. We kept the weights from the epoch with the lowest loss on the development set.

There were important mismatches in prosody during the recordings of the first and second session for the male speaker, with the second session leading to slower and more intelligible utterances. To better match the prosody of the male speaker during the second session, the models for the male speaker were further fine-tuned on the training data from the second recording for 5 epochs. Again, we kept the weights giving the lowest loss over the development set for the second recording session of the male speaker.

A ParallelWaveGAN vocoder was trained for each of the two speakers, using default hyper-parameters.

	Natural	Character	Phoneme
Male	4.34±0.16	3.93±0.24	3.73±0.21
Female	4.27±0.17	3.74±0.20	3.84±0.21

Table 4: Mean MOS score for each system, with 95% confidence intervals

	Natural	Character	Phoneme
Male	4.45±0.20	4.30±0.24	3.77±0.25
Female	4.54±0.15	4.25±0.23	4.04±0.26

Table 5: Mean speaker similarity score for each system, with 95% confidence intervals

6. Subjective Evaluation of Synthetic Speaker Similarity

6.1. Listening Tests Design

All 250 sentences of the test set were synthesized by each of the 4 systems (grapheme or phoneme based, both female and male voices). Out of the synthesized samples, 50 samples of length 3 to 5 seconds were randomly kept. Similarly, 50 samples of natural speech from the test set of both male and female speakers of length 3 to 5 seconds were chosen randomly. Those 300 samples were used as the basis for two listening tests, in order to evaluate the quality of the synthesis process and evaluate whether the original voices were reproduced faithfully. We contacted experienced Breton speaker to take part in the test. When asked to evaluate their skill on scale from 1 (Bad) to 5 (Excellent), all of the listeners answered with 4 or 5.

The first listening test was a MOS test including the 4 synthetic systems as well as the 2 natural voices. Listeners were asked to rate the overall quality of a randomly-chosen sample on a scale from 1 (Bad) to 5 (Excellent). Each listener was asked to rate 30 samples. 15 listeners participated in the test, so that every sample was rated at least once.

In the second listening test, a first sample was chosen randomly among the synthesized or natural samples. Then, a second sample of similar length and corresponding to a different prompt was chosen randomly among natural samples of the same speaker. Finally, the speakers were asked to rate how much they thought the two samples had been uttered by the same speaker. The rate was on a scale of 1 (Completely different speaker) to 5 (Identical speakers), with 3 being specified as "No opinion". Each listener was asked to rate 30 pairs. 11 speakers took part in the test, so that every sample was rated at least once.

6.2. Results

Table 4 shows the results of the listening test evaluating the overall quality of the speech synthesis systems. Listeners rated natural samples with a mean score of 4.34 and 4.27 for male and female voices respectively. These scores reflect the usual tendency of listeners to not give out perfect scores. The male synthetic voices received a mean score of 3.93 and 3.73 for the character and phoneme systems respectively. According to the annotated scale used during the listening test, this translates to a "good" quality. Similarly, the grapheme and phoneme-based systems for the female voice received 3.74 and 3.84, which translates to a "good" quality as well.

Table 5 shows the results of the listening test for the speaker similarity between synthetic (or natural) speech against natural

speech. Listeners rated the speaker similarity of natural samples with a mean score of 4.45 for the male samples, and 4.54 for the female samples. Since samples being compared by the listeners in a given pair did not correspond to the same prompt, those values suggest that the speakers were consistent during the recording of the dataset. The character-based systems for both male and female speaker were rated with fairly high-score, 4.30 and 4.25 respectively. Those values are not unexpected since we are using single-speaker systems. This suggests listeners agreed in saying that the character-based system was able to faithfully reproduce the voice and characteristics of the original speakers. The phoneme-based systems were rated lower than the character-based ones, with a score of 3.77 and 4.04 for the male and female systems respectively. Surprisingly, the difference is statistically significant for the male systems ($p < 0.05$ with a Mann-Whitney U test). This suggests that while using phoneme or grapheme inputs to train end-to-end TTS systems does not impact the quality of the overall synthesis process for Breton, it might impact the similarity with the original speaker.

7. Discussion

Further investigations are needed to confirm and understand the consequences of using grapheme or phoneme inputs on the speaker similarity for the Breton language.

The first investigation concerns the encoding of pause information. For the grapheme sequences, pauses were naturally encoded as the punctuation marks already present in the prompts. In that case the difference between short and long pauses are explicit and the models are able to reproduce them. For phoneme sequences, all types of pauses were encoded as a single special token. In that case, the models are not able to reproduce short and long pauses faithfully. This could lead to a difference in the prosody between natural and synthetic samples, explaining the lower results for the phoneme-based models.

The second investigation concerns the impact of the phonetisation process on the speaker similarity. As seen in Table 3, the phonetiser we developed is not perfect. The annotation of the phonetic sequences of the dataset used in the training of our models was done automatically, using our phonetisers. It is possible that training our systems on data containing phonetic annotation mistakes might have impacted the speaker similarity without lowering the overall quality of the samples. For example, erroneous predictions of *liaisons* by the phonetiser could lead to technically correct pronunciations and overall good quality samples, but a mismatch with the usual speaking pattern of the original speaker.

8. Conclusion

In this work, we tackled the problem of text-to-speech synthesis for an under-resourced language, Breton, with a particular focus on faithful reproduction of the original speaker voices. To solve this problem, we recorded a dataset with 20 hours of speech from two speakers highly pro-efficient in the target language. We developed tools to automatically annotate our speech corpus, such as a phonetiser. Due to the high amount of data we collected, we were able to use state of the art methods to perform text-to-speech synthesis, namely Tacotron2 and ParallelWaveGAN. This resulted in synthetic speech well-liked by native speakers during listening tests. Our focus on faithful speaker voice reproduction led us to compare grapheme and phoneme inputs with regard to speaker similarity. Surprisingly, grapheme-based models outperformed phoneme-based ones. Further investigations are needed to explain this gap in performance for the Breton language. In further works, we aim to take advantage of our recorded speakers being bilingual to add sentences containing only French words to the corpus. This would allow to perform code-switching to French. Furthermore, we plan to add new speakers to the database to get closer to our goal of preserving the Breton language.

9. Acknowledgements

This work has been funded by the public office of Breton language (*Office Public de la Langue Bretonne*) and done with the help of Skol Vreizh.

10. References

[1] Ailbhe Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl, "The abair initiative: Bringing spoken irish into the digital space," in *Proc. Interspeech 2017*, 08 2017.

[2] Michel Mermet, *Informatique et maîtrise de l'oral en maternelle bilingue breton-français: modèle de l'élève dans le dialogue enfant-ordinateur et ergonomie de la parole en breton.*, Ph.D. thesis, Université Rennes 2, 2006.

[3] Ander Corral, Igor Leturia, Aure Séguier, Michael Barret, Benaset Dazéas, Philippe Boula de Mareüil, and Nicolas Quint, "Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of gascon occitan," in *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop Language Resources and Evaluation Conference–Marseille–11–16 May 2020*. European Language Resources Association (ELRA), 2020.

[4] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[5] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[6] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[7] Menard Martial and Kadored Iwan and all., *Geriadur brezhoneg*, An Here, Plougastel-Daoulas, 2001.

[8] Favereau, Francis, *Geriadurig ar brezhoneg a-vremañ*, Skol Vreizh, Morlaix, 2015.

[9] Le Ruyet Jean-Claude, *Bien prononcer le breton d'aujourd'hui ; les liaisons*, Skol Vreizh, Morlaix, 2012.

[10] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda,

Kazuya Takeda, Yu Zhang, and Xu Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.