Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 53–60
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Neural Text-to-Speech Synthesis for an Under-Resourced Language in a Diglossic Environment: the Case of Gascon Occitan

Ander Corral[1], Igor Leturia[1], Aure Séguier[2], Michäel Barret[2], Benaset Dazéas[2], Philippe Boula de Mareüil[3], Nicolas Quint[4]

[1] Elhuyar Foundation
{a.corral, i.leturia}@elhuyar.eus
[2] Lo Congrès permanent de la lenga occitana
{a.seguier, m.barret, b.dazeas}@locongres.org
[3] Université Paris-Saclay, CNRS, LIMSI
philippe.boula.de.mareuil@limsi.fr
[4] Laboratoire Langage, Langues et Cultures d'Afrique (LLACAN - UMR 8135 - CNRS/INALCO/USPC)
nicolas.quint@cnrs.fr

Abstract

Occitan is a minority language spoken in Southern France, some Alpine Valleys of Italy, and the Val d'Aran in Spain, which only very recently started developing language and speech technologies. This paper describes the first project for designing a Text-to-Speech synthesis system for one of its main regional varieties, namely Gascon. We used a state-of-the-art deep neural network approach, the Tacotron2-WaveGlow system. However, we faced two additional difficulties or challenges: on the one hand, we wanted to test if it was possible to obtain good quality results with fewer recording hours than is usually reported for such systems; on the other hand, we needed to achieve a standard, non-Occitan pronunciation of French proper names, therefore we needed to record French words and test phoneme-based approaches. The evaluation carried out over the various developed systems and approaches shows promising results with near production-ready quality. It has also allowed us to detect the phenomena for which some flaws or fall of quality occur, pointing at the direction of future work to improve the quality of the actual system and for new systems for other language varieties and voices.

Keywords: TTS, Occitan, Gascon, Tacotron 2, WaveGlow

1. Introduction

1.1. The Occitan Language and the Gascon Variety

Occitan is a Romance language spoken in three states of the European Union (France, Spain and Italy) on an area of about 180,000 km^2. The Occitan language is co-official with Catalan and Spanish in Val d'Aran (in the Pyrenees). Although it has no public recognition elsewhere, several French territorial collectivities support it and it is one of the languages considered by the linguistic minorities protection law in Italy. As there are no officially imposed standards, Occitan may be described as a fairly dialectalised language. It is traditionally divided into six major varieties (Bec, 1986; Quint, 2014): Gascon, Languedocian, Provençal, Limousin, Auvergnat and Vivaro-Alpine. Gascon, the South-Western variety, displays many specific features (on both phonological and morphosyntactic levels), partly due to the Aquitanic substrate shared with Euskara (Basque). Perceptually, the presence of the [h] sound is particularly salient in the phonemic inventory: /p b t d k ɡ n ɲ m f s z ʃ ʒ h r l ʎ i ɥ j w e ɛ a ɔ y u/.

1.2. Language Technologies for Occitan

Occitan being in a minorised language situation, it is also one of the many "under-resourced" or "poorly endowed" languages in the area of Natural Language Processing (NLP). However, NLP is a prioritary issue for its development and its diffusion. Hence, there was a need to develop a strategy according to the existing means, with planned goals as described in the Digital Language Survival Kit (Ceberio et al., 2018). The Occitan language benefits from some resources: text corpora such as Batelòc (Bras and Vergez-Couret, 2008) or Lo Congrès' one; online dictionaries or lexica such as dicod'Òc, tèrm'Òc, Lofłòc (Bras et al., 2017); a verb conjugator (vèrb'Òc), among others. Some tools have also been tested on Occitan: Talismane PoS tagger (Urieli, 2013), used within the framework of the ANR Restaure program (Bernhard et al., 2019), a spell checker, a predictive keyboard (Congrès, 2018a; Congrès, 2018b) and an automatic translator on the Apertium platform (Apertium, 2016).

These resources are mentioned in the Roadmap for Occitan Digital Develoment (Gurrutxaga and Leturia, 2014), a document for the planification of the development of Occitan NLP resources based on the Meta-Net method (Rehm and Uszkoreit, 2012). An Occitan text-to-speech (TTS) system is planned for 2019, also mentioned in the Inventory of linguistic resources for the languages of France (Leixa et al., 2014). In this report produced by the French Ministry of Culture (DGLFLF), the resources for Occitan are described as weak.

2. A Text-To-Speech System for Gascon

The recomendations for 2019 of the aforementioned Roadmap for Occitan Digital Develoment included building a Text-To-Speech (TTS) system. In accordance with

these recommendations, work aiming at developing TTS technologies for Occitan took off. As a first approach, we decided to perform some experiments and evaluations with one variety and one voice. Once the desired results would be achieved (and therefore, the technology chosen, the process mastered and the sizes and features of the training datasets known), we would then expand the same methodology to other voices and varieties.

This first work was carried out for the Gascon variety with a female voice. However, there were some restrictions or preconditions to take into account in the experiments, which we will describe in the following subsections.

2.1. Deep Neural Network Technology

In recent years there has been a great change in the TTS area, whereby systems have progressively shifted from parametric or concatenative methods to deep neural network based ones. Naturally, this was due to the great improvement in quality and naturalness obtained by the latter technique.

The shift started with Google publishing the first WaveNet paper (Oord et al., 2016), a deep neural network model to implement the last steps of a TTS system, that is to say, the vocoder and acoustic modeling part (they produce the speech waves out of linguistic or acoustic features). WaveNet largely surpassed existing systems in terms of naturalness and quality. The paper was later followed by another one on Parallel WaveNet (Oord et al., 2018), a faster implementation of the same principle.

Tacotron (Wang et al., 2017) accelerated the shift. It was another neural network model to get spectrograms directly from text, instead of having the usual multiple steps chain (text analysis, duration modeling, acoustic feature modeling...) where errors tend to accumulate. Combined with a simple waveform synthesis technique, Tacotron outperformed production parametric systems in terms of naturalness. Tacotron 2 (Shen et al., 2018) put the final nail, proving that they could obtain a naturalness score almost as good as professionally recorded speech by generating mel spectrograms using the Tacotron approach and subsequently generating the speech waveform using a WaveNet model.

Also most of the other TTS systems that have come into scene in recent years, like Deep Voice (Arik et al., 2017) or Char2Wav (Sotelo et al., 2017) are also using deep learning approaches and improving the quality of previously used methods.

In view of the results of the Tacotron-WaveNet method, free software implementations of Tacotron and WaveNet have arisen, to allow researchers to perform experiments and developers to produce systems for other languages or situations. One of the most prominent is NVIDIA's, who have published under a free license a Tacotron 2 implementation (NVIDIA, 2018a) and WaveGlow (NVIDIA, 2018b), a system combining WaveNet and Glow (Kingma and Dhariwal, 2018) which, according to a paper they released (Prenger et al., 2019), delivers audio quality as good as the best publicly available WaveNet implementation. For these reasons, we chose NVIDIA's Tacotron 2 and WaveGlow combination as the software to perform our experiments with Gascon and, if successful, to put into production TTS systems for Occitan in general.

2.2. Few Recording Hours

Occitan being an under resourced language, audio recordings such as those needed for training TTS systems (good quality, one speaker, transcribed, aligned and in large quantities) are not available for Occitan. This means that recordings had to be made specifically for the project. And taking into account that TTS systems for many voices and varieties are planned to be developed in the future, we needed to adjust the recording hours to the minimum required. But since the amount of recording hours needed to obtain good quality cannot be known beforehand, we decided to start with a small amount of hours and evaluate the results obtained, and then make more recordings afterwards if necessary.

However, we needed a starting point of reference for the amount of recording hours. The Google experiments mentioned in the Tacotron and WaveNet papers use proprietary training datasets of at least 25 hours for each language (English and Chinese). The system mentioned in the WaveGlow paper uses the free LJ Speech dataset (Ito, 2017), which also represents 24 hours of audiobooks recorded in English. As we have already stated, our goal was to do it with much fewer hours if possible.

To our knowledge, the majority of research work using the deep neural Tacotron-WaveNet approach was based on the above referred training datasets, and there is not much work mentioning other datasets or languages. (Yasuda et al., 2019) have developed a system for Japanese, but they use a 47 hour dataset.

Some other works like Latorre et al. (2019) prove that, in the absence of many recording hours from a single speaker, a similar or better quality can be achieved with few hours from many speakers; but this does not help us, since there is not a corpus of this kind for Occitan anyway. And others like Tits et al. (2020) focus on developing new voices with few recordings, but this requires having already a multi-speaker TTS system, which raises the same issue regarding Occitan. Finally, Chen et al. (2019) explore the development of a new voice with few hours using a TTS system developed for another language, with promising, albeit experimental, results.

One reference that could be of use for our work is the one described by Choi et al. (2018), where they made experiments with around 9 hours data. However, the Mean Opinion Score (MOS) they achieved is far below the levels reported in the rest of the above papers, which is not very promising. On the other hand, Podsiadło and Ungureanu (2018) prove that a quality dataset (phonetically balanced, professional quality recordings) of 10 hours can achieve almost as good MOS scores as a 23 hour dataset, and Liu et al. (2019) also achieve good MOS scores with 8 hours. So we decided to start with a similar number of hours, which was in principle affordable to us, as a first experiment, which we would forcibly enlarge if the experiment did not achieve good results.

2.3. Standard Pronunciation of French Proper Names

Due to the sociolinguistic situation of the region where Occitan is spoken, where there is a remarkable diglossia of Occitan with respect to French, many French words are included in Occitan oral and text production (notably proper names of people, streets, places, brands, titles, etc.). These words are usually pronounced as in French, and if they are pronounced following the traditional phonological rules of Occitan, the result is incomprehensible for the speakers themselves. Therefore, if an Occitan TTS would not take this into account and pronounce French names like "Beauvais" or "Jeanssins" with the Occitan pronunciation as [be/aw/'bajs] and [ʒe/an/'sis], rather than [/bo/'vɛ] and [/ʒã/'sɛs̃], it would not be good for practical use with real texts.

This same problem had been detected in Iparrahotsa (Navas et al., 2014), a TTS system for the Navarro-Lapurdian variety of Basque, spoken in the French part of the Basque Country. The project is still ongoing, precisely due to that problem.

For this reason, in order to be properly developed, the TTS system for Occitan had to correctly pronounce French proper names following the French (vs. Occitan) phonological system, so we also needed to record French words, to test phoneme-based approaches and to develop language detection and text-to-phoneme conversion tools.

In the aforementioned work by Yasuda et al. (2019), they perform some experiments using phonemes as input. Other systems using phonemes or mixed input are mentioned in Kastner et al. (2019).

3. Experiments

3.1. Training Dataset

3.1.1. Text Corpus

We designed a relatively small Occitan corpus made up of literary works, press articles and Wikipedia pages. In order to maximise the recording time in terms of phoneme diversity, we sorted all its sentences according to a score indicating the *number of unique diphones/total number of diphones* ratio and the proximity to the average sentences length, inspired by the "Unique Unit Coverage Score" introduced by Arora et al. (2014).

The corpus aimed at including all diphones which might occur in an Occitan conversation mixed with French words: every possible combination of Occitan and French phonemes was therefore taken into account. We first picked up in the Occitan corpus all sentences showing a diphone that did not appear in the sentences already selected. Then, we manually added sentences containing diphones which did not occur in the corpus.

In addition, we used a list of phonetised French given names and family names (Boula de Mareüil et al., 2005) as an exception database (see Section 3.2.). Those proper names were then combined automatically to get a corpus with all diphones combining a French phoneme and a French or an Occitan phoneme (by "French phonemes" we mean the 10 vowels and consonants which do not belong to the standard Occitan phoneme inventory, such as nasal vowels). We manually created sentences including proper names with diphones which did not occur in the resulting corpus. Sentences including diphones with the Spanish *jota* (/*x*/) (which does not belong to the phoneme inventories of Occitan and French) were added in order to account for Spanish loans, whose frequency is particularly significant in written texts produced in Aranese Occitan (a Gascon variety spoken alongside Spanish in the Val d'Aran).

As a result, we obtained a first "mandatory" corpus containing all possible diphones formed with Occitan and French phonemes (as well as the *jota*). Also, we manually added sentences which are likely to appear in systems using TTS synthesis, such as GPS or public transport (e.g., "Turn left"). Finally, we picked up sentences in our sorted Occitan corpus until we obtained a total of 100,000 words. We obtained a corpus of approximately 13,600 sentences: about 10,900 entirely Occitan sentences, around 2,700 sentences with French phonemes, 43 sentences with the Spanish jota. There were around 1,300 exclamatory sentences and 700 interrogative sentences.

3.1.2. Audio Recordings

We looked for Occitan speakers suited for the task, that is, speaking for hours in a "neutral tone", and we chose to hire a female radio news announcer. We recorded her in a studio usually dedicated to movies dubbing. Therefore, the sound engineer present in the studio was used to working with spoken material. The recordings lasted six days, with an average of seven hours work per day. Sentences were projected on a large screen about seven meters away from the speaker, in order to elicit a voice similar to that of someone speaking from a certain distance. After cleaning, we obtained a total of almost 7 hours of speech (without counting the pauses between sentences, which would otherwise add a few hours). We were, however, not able to record the whole written corpus.

We had to deal again with the internal great variability of the Occitan language, even within the Gascon domain. We wanted our TTS system to fit the regional standard as much as possible. However, we soon realized that it was impossible to have our speaker speak during hours with some pronunciations that were unnatural to her (e.g.: pronounce the j /ʒ/ instead of /*j*/). These pronunciation variations were rather few and were not pointed out during the quantitative evaluation. Still, we had to settle for compromises in favour of a less standard Occitan (Gascon): the result is still a reasonably standard accent, easily understandable by most Occitan speakers. Moreover, after the recording sessions, we spent a lot of time in post-processing to make adjustments. In particular, many sentences of our corpus were not written in the speaker's subvariety, and she pronounced them following her own speech habits. It was thus necessary to correct a number of written sentences to make them fit the recorded pronunciation. Also, we took advantage of this opportunity to correct the mispronounced words.

The size of the corpus is 6:52 hours, out of which 5:46 contained only Occitan words and another 1:06 hours contained sentences including French (or other languages) words. A total of 23 minutes were recordings of interrogative sentences and 36 minutes of exclamatory sentences.

The mean sentence length is around 2.5 seconds, with a standard deviation of around 1.8. However, some of the sentences were much longer. Those exceeding 10 seconds triggered memory errors on training, so we had to remove them (they accounted for fewer than 1% of the sentences).

3.2. Linguistic Tools

We created an expansion tool which cleans the input and expands Arabic numerals and most Roman numerals, e-mail addresses, website URLs, phone numbers, dates, hours, measurement units, currencies, some acronyms and abbreviations. A rule-based grapheme-to-phoneme conversion tool was developed, containing about 230 hand-written rules. It uses an exception base (including thousands of French proper names) as well as a syllabification tool composed of 40 rules, for lexical stress assignment. Occitan is a language in which stress is distinctive, and stress location may be predicted in most cases even in the absence of part-of-speech tagging — a component which has not yet been designed.

3.3. TTS Systems

3.3.1. The Tacotron Part

We devised three setups for our experiments, all of them using the NVIDIA's Tacotron 2 and WaveGlow above mentioned systems. The first one (hereinafter OccTxt) was trained with those sentences that only contain Occitan words, in their character-based version (with numbers, acronyms, etc. expanded using the tool described above). The second one (hereinafter OccPho) was also trained only with Occitan sentences, but these were converted to phonemes by the tool mentioned in the previous subsection. Finally, a third setup (AllPho) was prepared using all available sentences (including French proper names) converted to phonemes (the French words being transcribed according to their standard French pronunciation).

We are aware that Tacotron 2 can accept a mixed character-phoneme notation as an entry (the code includes the example "Turn left on {HH AW1 S S T AH0 N} Street") and that for the third setup we could have used such a notation with Occitan text in characters and only French words as phonemes. However, since the grapheme-to-phoneme conversion tool we developed could convert all the text to phonemes, we considered this to be a better option, because it reduced the symbol set (some of the French phonemes also exist in Occitan) and so we could expect better results. Tacotron 2 includes a list of the accepted letters and symbols (which are the English ones). We changed this list to also accept Occitan diacritics in the first setup, Occitan phonemes in the second and Occitan and French phonemes in the third. The list of characters and phonemes was obtained from the training corpus which, depending on the setup, was passed through the phonemizer or not.

OccTxt and OccPho were designed to test if an acceptable quality could be achieved in Occitan with a relatively small number of recording hours, although we did not expect them to work well with French proper names. AllPho would serve both to check (i) if French words were well pronounced, and (ii) if the inclusion of French words into the training set would impact the pronunciation of the Occitan words with respect to the other setups.

In all cases, we used as a starting point the models NVIDIA trained for English with their Tacotron 2 implementation using the LJ Speech dataset, which are both downloadable from their GitHub page (NVIDIA, 2017a). As the authors of these systems explain, "training using a pre-trained model can lead to faster convergence", which proved to be the case: in some experiments we carried out there was no noticeable difference between the results of a system trained on a random state and a system trained on the English model, with a much smaller training time for the latter. For the training phase, default parameters were used, with no hyperparameter optimization. Models were trained until no further improvement was obtained on the validation data.

3.3.2. The WaveGlow Part

Whatever the results might be, we did not expect a quality comparable to what was reported in the original Tacotron 2 - WaveNet paper, because of the much smaller amount of available hours of recordings. But it was interesting to see (i) if the effect of this reduced corpus could be more significant at the level of the production of mel spectrograms from text (the Tacotron part) or at the level of the production of the audio wave from the mel spectrograms (the WaveGlow part), and (ii) if we were able to improve the results by somehow intervening in one of those steps.

We observed that using a WaveGlow model trained for English with more hours -precisely, the one reported in the WaveGlow paper, trained with the 24 hours LJ Speech dataset and also downloadable from its GitHub page (NVIDIA, 2017b)- could also be used as vocoder with a relatively good quality for Occitan. This result may seem curious at first, but was in some way logical: mel spectrograms are nothing but frequency-based representations of a sound wave, and all that a system like WaveGlow or WaveNet does is learning to produce an audio wave from a mel spectrogram; therefore, if a system learns from a large dataset where many frequencies are represented, it should be able to decode mel spectrograms of other voices and languages quite efficiently, unless the new language and voice contain many frequencies both new and unknown to the system. In our case, both the LJ Speech dataset and the Occitan training dataset were female voices (and so supposedly relatively near from each other in the frequency spectrum), and the English model seemed to fit well for the Occitan spectrograms.

Therefore, for the WaveGlow step, in additon to the models trained on the Occitan audios (henceforth OccWav), we also tested a model produced from the English LJ Speech dataset (henceforth EngWav).

4. Evaluation and Results

For the evaluation, a corpus of 100 sentences was prepared, containing examples of all the phenomena we wanted to test, with the following distribution: 10 exclamations, 20 interrogations, 15 with rare diphones, 20 with at least one French noun, and the remaining 35 being affirmative sentences with only Occitan words. These sentences were all

recorded by the speaker of the training set.

4.1. The Tacotron Part

In a first phase, evaluators were presented with these sentences both in (i) their recorded and (ii) their synthesized version produced using in turn the three Tacotron setups (OccTxt, OccPho and AllPho) with the EngWav model in the WaveGlow part, i.e. 400 audios, in a blind random way. For each of these, the evaluators were required to evaluate three points:

- If the sentence was correctly pronounced (in a scale from 1 to 5), which would allow us to detect pronunciation errors in French proper names, question intonations, rare diphones, etc.
- If the sentence was fluid and natural (in a scale from 1 to 5), so that we could have a MOS (Mean Opinion Score) of the voices' naturalness.
- If there was a major technical problem such as truncated sentence, blank, gibberish... (yes/no rating), because we observed that such things sometimes happened due to the small amount of recorded hours, the length of some sentences or other reasons.

The evaluations of the first phase were done by 8 Gascon-speakers over a period of ten days, which represents a total of 3,200 (= 8 x 400) evaluated sentences. The evaluators are professionals working with Occitan language in many fields (teaching, administration, linguistics, translation...). They were not familiar with TTS systems, but had occasionally heard synthesized speech in French (GPS, public transport...).
The MOS obtained for the pronunciation correctness by the three systems and compared to the human recordings can be seen in Figure 1. The system that scores best is AllPho, obtaining a score of 4.2, whereas the human recordings obtained 4.9. Besides, it has a difference of at least 0.3 with respect to the other two systems, which was somehow expected, since the other systems lacked the phoneme conversion or the recordings for French, which could only make them score lower in the sentences containing French proper names.

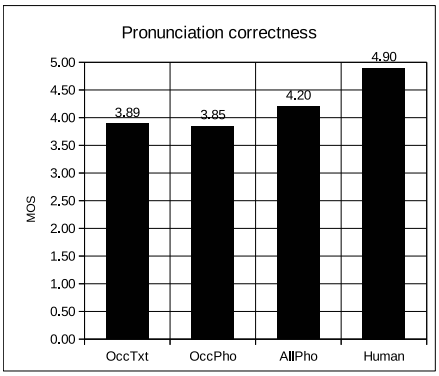

Figure 1: MOS for pronunciation correctness.

In fact, if we look at the scores obtained by each sentence type (Table 1), we can see that, although the AllPho setup scores best for every type of sentence, the improvement obtained is much larger for the sentences with French nouns (almost 0.5 points).

Sentence	System			Human
	OccTxt	OccPho	AllPho	
Normal	3.96	3.93	**4.23**	4.91
French	3.47	3.64	**4.08**	4.81
Interrogative	3.91	3.81	**4.10**	4.95
Exclamatory	4.53	4.04	**4.65**	4.93
Rare diphones	3.85	3.87	**4.17**	4.90
Average	3.89	3.85	**4.20**	4.90

Table 1: MOS for pronunciation correctness for each type of sentence.

The AllPho system is also the best one regarding fluidity and naturalness, as is shown in Figure 2. This can be the result of having 15% extra-time of recordings for training, because the sentences with French words also include Occitan text and therefore a larger audio corpus of Occitan was used. The score obtained can be considered as satisfactory, taking into account the relatively small number of training hours used.

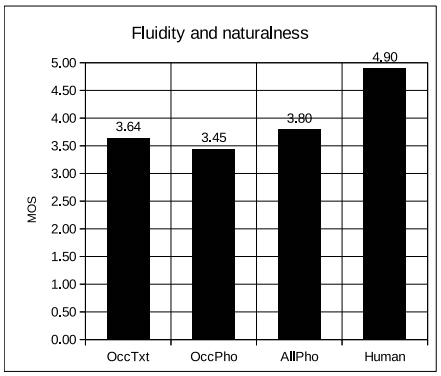

Figure 2: MOS for fluidity and naturalness.

Table 2 shows us that AllPho scores best in all types of sentences. This is a very important positive result, because it means that the inclusion of French phonemes in the recordings and the fact of basing the TTS system on phonemes instead of text characters does not impact negatively the voice quality; rather, it slightly improves it.

Sentence	System			Human
	OccTxt	OccPho	AllPho	
Normal	3.85	3.60	**3.93**	4.92
French	3.25	3.26	**3.63**	4.83
Interrogative	3.56	3.33	**3.60**	4.96
Exclamatory	4.15	3.60	**4.21**	4.88
Rare diphones	3.48	3.39	**3.68**	4.89
Average	3.64	3.45	**3.80**	4.90

Table 2: MOS for fluidity and naturalness for each type of sentence.

We can also observe in Tables 1 and 2 that exclamatory sentences obtain the highest score both in correctness and naturalness in all setups, with a significant difference with respect to the other types of sentence especially in the AllPho system. Interrogative sentences, on the other hand, get the lowest score in naturalness and the second lowest in correctness (very close to the lowest one, i.e. sentences containing French names) in the AllPho setup, with a difference of more than 0.5 points with respect to exclamatory sentences. This may be due to the fact that there are 56% more recordings of exclamatory sentences than interrogative sentences in the corpus (36 min. vs. 23 min., see Section 3.1.2.). Recording some extra interrogative sentences (just 13 more minutes) in order to have the same amount of recordings as available for exclamatory sentences might be a way to improve the synthesis of interrogative sentences. However, we cannot exclude that other factors (such as pragmatic saliency or cognitive aspects) may also account for this difference between the two sentence types.

Finally, regarding the third question (if the sentence presented some major technical error or problem), as can be seen in Table 3, there does not seem to be any significant difference among the three systems, although AllPho seems more reliable. All three systems seem to meet with similar difficulties in correctly synthesizing the same specific sentences.

Sentence	System			Human
	OccTxt	OccPho	AllPho	
Normal	49	45	**46**	4
French	36	38	**30**	3
Interrogative	24	25	**26**	3
Exclamatory	3	8	**3**	1
Rare diphones	19	16	**18**	1
Average	131	132	**123**	12

Table 3: Major technical problems or errors for each type of sentence.

4.2. The WaveGlow Part

In a second phase, we prepared an evaluation that would play the 100 sentences synthesized by the system that scored best in the first phase (AllPho) with the EngWav and OccWav WaveGlow vocoders, that is, 200 audios, again in blind and random conditions. In this case, the evaluators only had to evaluate the fluidity and naturalness and give it a score on a 1 to 5 scale. However, listening to just some few sentences was enough for the evaluators to clearly identify which ones were produced with the EngWav vocoder used in the first phase or with the new OccWav one. The evaluators reported that EngWav was by far better and more natural, and saw no point in going on with a tiring and costly evaluation.

4.3. Qualitative Evaluation

After finishing the quantitative evaluation, the evaluators were asked two questions about their qualitative impression on the systems developed:

- What was their global impression on the sentences they had heard.
- The type of technical or quality errors or problems they had encountered.

Globally speaking, the evaluators find that the synthesized sentences (they were usually distinguishable from the human recorded ones) were easy to understand and of good quality. Taking into account the fact that the evaluators did not know which system had produced each sentence, if they thought the synthesized sentences had a good average quality, we can suppose that the system which got the highest scores (AllPho) would be considered by the same evaluators still better.

Regarding the errors or problems, the evaluators mentioned occasional silences or missing words, noise (which they qualified as whistling, blowing, metallic..), artificiality, intonation problems and difficulties with specific words.

A more detailed analysis of the sentences marked by the evaluators as problematic enabled us to see the causes of some of these problems and devise possible solutions for a production system. For example, some of the silence and intonation problems are due to the fact that the systems implemented get confused when producing very long sentences (as mentioned earlier, we removed sentences longer than 10 seconds from the training datasets because they resulted in memory errors); therefore, to solve this problem, we divided long sentences at points where a comma was found, synthesized them as separate sentences and concatenated them; but this produced too long silences and intonation falls at commas (similar to those at the end of a sentence). We believe that if we cut some long sentences at commas from the training dataset, this problem can be relieved. This work is yet to be done. Likewise, it seems that by applying a filter to reduce the trebles by 10 dB, the whistling or metallic effect is reduced, although we have not yet tested this effect formally.

5. Conclusions

With the objective of obtaining a neural state-of-the-art TTS for Gascon Occitan with relatively few recording hours which would pronounce French proper names in a standard way, we have developed and evaluated different systems, all based on the NVIDIA Tacotron 2 - WaveGlow software, some of them text-based and others phoneme-based, some of them including recordings of French words and some not. The system based on phonemes which included the recordings of French words obtained a MOS of 4 out of 5 for correctness and naturalness, and evaluators and language experts consider it to be of a near production-ready quality.

Moreover, the evaluation results were useful to show which sentence types or phenomena may need improvements or adjustments and which types of sentence produce some major problems. Basing ourselves on these results, we will be able to decide if further recordings of some kinds of sentences must be done or if we should choose other technical solutions in order to solve the various problems encountered, with a view to putting into production a TTS system for Gascon Occitan and developing systems for other voices

and other Occitan varieties such as Languedocian. The idea is to put these systems into production and to make them available for interested users, organizations and companies in the short term. The MOS of these systems might not be as high as that of systems trained with more hours, but it was considered to be of a satisfactory quality, especially if we take into account the fact that there are no TTS systems currently available for Occitan. At any rate, this Occitan TTS system will probably be better than many other systems produced for other languages with older technologies. However, it is important to note that the system we have chosen to put into production makes use of a vocoder model produced for English, because it sounds much better to the evaluators than the model trained specifically for Occitan. This means that the recording hours used in the project might not always be sufficient to train a WaveGlow system with a production-ready quality. Here a model trained for another language with more recording hours was useful, but it might not always be the case. For example, the model may not be useful for correctly synthesizing male voices due to a different frequency spectrum, since it was trained using recordings of a female speaker. Also, to our knowledge, there are no free WaveGlow models trained with large datasets of male voices. Thus, if we are faced with this problem in the future, we will have to try and train a model using free recorded datasets of male voices (if available), or maybe endeavour to expand our own set of recordings.

6. Acknowledgements

The research carried out in this project is part of the project "LINGUATEC: Development of cross-border cooperation and knowledge transfer in language technologies" (POCTEFA EFA227/16, ERDF), funded by the Ministry of Economy and Competitiveness of Spain and the European Regional Development Fund (ERDF).

7. Bibliographical References

Apertium. (2016). Apertium translation pair for Occitan and French. https://github.com/apertium/apertium-oci-fra.

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., and Shoeybi, M. (2017). Deep Voice: Real-time Neural Text-to-Speech. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sidney, Australia, July.

Arora, K., Arora, S., Verma, K., and Agrawal, S. S. (2014). Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages. In *Proceedings of 8th International Conference on Spoken Language Processing (Interspeech 2004 - ICSLP)*, Jeju, Korea, October.

Bec, P. (1986). *La langue occitane*. Number 1059 in Que sais-je? Presses universitaires de France, Paris, 5th edition.

Bernhard, D., Bras, M., Erhart, P., Ligozat, A.-L., and Vergez-Couret, M. (2019). Language Technologies for Regional Languages of France: The RESTAURE Project. In *Proceedings of International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*, Paris, France, December.

Boula de Mareüil, P., d'Alessandro, C., Bailly, G., Béchet, F., Garcia, M.-N., Morel, M., Prudon, R., and Véronis, J. (2005). Evaluating the Pronunciation of Proper Names by Four French Grapheme-to-Phoneme Converters. In *Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 1251–1254, Lisbon, Portugal, September.

Bras, M. and Vergez-Couret, M. (2008). BaTelÒc: A Text Base for the Occitan Language. *Language Documentation & Conservation*, Special Publication No. 9(Language Documentation and Conservation in Europe):133–149.

Bras, M., Vergez-Couret, M., Hathout, N., Sibille, J., Séguier, A., and Dazéas, B. (2017). Lofl̀oc, lexic obèrt flechit occitan. In *Proceedings of the XIIn Congrès de l'Associacion internacionala d'estudis occitans*, Albi, France, July.

Ceberio, K., Gurrutxaga, A., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality. Technical report, The Digital Language Diversity Project.

Chen, Y.-J., Tu, T., Yeh, C.-c., and Lee, H.-Y. (2019). End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proceedings of Interspeech 2019*, pages 2075–2079, September.

Choi, Y., Jung, Y., Kim, Y., Suh, Y., and Kim, H. (2018). An end-to-end synthesis method for Korean text-to-speech systems. *Phonetics and Speech Sciences*, 10(1):39–48, March.

Congrès, L. (2018a). Occitan gascon pack for AnySoftKeyboard. https://play.google.com/store/apps/details?id=com.anysoftkeyboard.languagepack.gascon.

Congrès, L. (2018b). Occitan lengadocian pack for AnySoftKeyboard. https://play.google.com/store/apps/details?id=com.anysoftkeyboard.languagepack.lengadoc.

Gurrutxaga, A. and Leturia, I. (2014). Diagnostic et feuille de route pour le développement numérique de la langue occitane : 2015-2019. Technical report, Elhuyar Foundation, Media.kom, November.

Kastner, K., Santos, J. F., Bengio, Y., and Courville, A. (2019). Representation Mixing for TTS Synthesis. In *Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5906–5910, Brighton, United Kingdom, May. IEEE.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative Flow with Invertible 1x1 Convolutions. *arXiv:1807.03039 [cs, stat]*, July. arXiv: 1807.03039.

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Klimkov, V. (2019). Effect of Data Reduction on Sequence-to-sequence Neu-

ral TTS. In *Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7075–7079, Brighton, United Kingdom, May.

Leixa, J., Mapelli, V., and Choukri, K. (2014). Inventaire des ressources linguistiques des langues de France. Technical Report ELDA/DGLFLF-2013A, ELDA, November.

Liu, B., Chen, Y., Yin, H., Li, Y., Lei, X., and Xie, L. (2019). The Mobvoi Text-To-Speech System for Blizzard Challenge 2019. In *Proceedings of The 10th ISCA Speech Synthesis Workshop (SSW10)*, Vienna, Austria, September.

Navas, E., Hernaez, I., Erro, D., Salaberria, J., Oyharçabal, B., and Padilla, M. (2014). Developing a Basque TTS for the Navarro-Lapurdian Dialect. In *Proceedings of IberSPEECH 2014*, volume 8854, pages 11–20, Las Palmas de Gran Canaria, Spain.

NVIDIA. (2018a). NVIDIA/tacotron2, May. https://github.com/NVIDIA/tacotron2.

NVIDIA. (2018b). NVIDIA/waveglow, November. https://github.com/NVIDIA/waveglow.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September. arXiv: 1609.03499.

Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., and Hassabis, D. (2018). Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In Jennifer Dy et al., editors, *Proceedings of 35th International Conference on Machine Learning*, volume 80, pages 3918–2926, Stockholm, Sweden, July. arXiv: 1711.10433.

Podsiadło, M. and Ungureanu, V. (2018). Experiments with Training Corpora for Statistical Text-to-speech Systems. In *Proceedings of Interspeech 2018*, pages 2002–2006, September.

Prenger, R., Valle, R., and Catanzaro, B. (2019). Waveglow: A Flow-based Generative Network for Speech Synthesis. In *Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, Brighton, United Kingdom, May.

Quint, N. (2014). *L'Occitan*. Assimil, Chennevières sur Marne, France.

Georg Rehm et al., editors. (2012). *META-NET White Paper Series: Europe's Languages in the digital age*. Springer, Heidelberg, Germany.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, Calgary, Canada, April.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. (2017). Char2Wav: End-to-End Speech Synthesis. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, April.

Tits, N., El Haddad, K., and Dutoit, T. (2020). Exploring Transfer Learning for Low Resource Emotional TTS. In Yaxin Bi, et al., editors, *Intelligent Systems and Applications*, volume 1037, pages 52–60. Springer International Publishing.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université Toulouse 2 Le Mirail, Toulouse, France, December.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. In *Proceedings of Interspeech 2017*, pages 4006–4010, Stockholm, Sweden, August. ISCA.

Yasuda, Y., Wang, X., Takaki, S., and Yamagishi, J. (2019). Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In *Proceedings of ICASSP 2019*, pages 6905–6909, Brighton, UK, May. arXiv: 1810.11960.

8. Language Resource References

Ito, K. (2017). The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset.

NVIDIA. (2017a). Tacotron 2 model weights pre-trained on the LJ Speech dataset. https://drive.google.com/file/d/1c5ZTuT7J08wLUoVZ2KkUs_VdZuJ86ZqA/view.

NVIDIA. (2017b). WaveGlow model weights pre-trained on the LJ Speech dataset. https://ngc.nvidia.com/catalog/models/nvidia:waveglow_ljs_256channels.