# The CSTR entry to the Blizzard Challenge 2021

*Dan Wells, Pilar Oplustil-Gallegos, Simon King*

The Centre for Speech Technology Research, University of Edinburgh, UK

{dan.wells, P.S.Oplustil-Gallegos, Simon.King}@ed.ac.uk

## Abstract

We describe the text-to-speech (TTS) system submitted from The Centre for Speech Technology Research at the University of Edinburgh to the Blizzard Challenge 2021. We participated in the spoke task to build a voice for Peninsular Spanish, where test utterances contain a small number of English words. Our system is trained from monolingual data in Spanish and English, including some Spanish-accented English and Spanish utterances containing English words, but without explicit supervision for these aspects. Input texts are represented using phonological feature vectors to encourage parameter sharing between the two languages despite different phoneme inventories. When synthesizing test utterances, we perform automatic language identification to provide word-level language embeddings and apply pronunciation nativization rules to any detected English words to bring them closer to native Spanish phonology. In addition to the results of the main Blizzard Challenge evaluation, we present analysis of the impact of nativization strategy on listener preferences, which may be relevant for evaluation of code-switching TTS in general.

**Index Terms**: Blizzard Challenge, speech synthesis, code-switching

## 1. Introduction

We submitted an entry to the spoke task of the Blizzard Challenge 2021, synthesizing Peninsular Spanish speech for texts containing a small number of English words in each sentence.

The pronunciation of English words by speakers of Peninsular Spanish is characterized by nativization into Spanish phonology [1]. This means that, for example, phonemes that occur only in English are adjusted to fit the phonological inventory of Spanish. English has always been a common source of loan words for every Spanish dialect, but even more so in recent years due to the influence of Anglophone social networks, technology and mass media, which has resulted in Spanish speakers' frequent use of everyday English loans such as "sorry, gym, spoiler, OK, email, online" [2].

The problem for synthesizing Spanish utterances which include English loan words is that predicting how native Spanish speakers would pronounce those loan words is subject to cultural and social variables, and in fact there is no unique right answer in any given case. For example, it is unlikely that even a bilingual speaker of Spanish and English would adopt a native pronunciation for an English loan word in a Spanish sentence. The challenge is therefore to nativize these words to a degree which is appropriate for the target listeners, here speakers of Peninsular Spanish. Although previous work has found that data-driven approaches outperform rule-based methods [1], in this paper we use a rule-based approach given the lack of available data and the time to collect it.

We also need to consider how to represent input texts containing a mixture of two languages. Previous work has shown phonological feature vectors to be a viable input representation for zero-shot synthesis of unseen phonemes from a second language in code-switched text [3], as well as for multilingual TTS including low-resource languages [4]. Individual phonological features (PFs) represent distinctive articulatory aspects of phonemes such as tongue position, degree of closure and voicing [5], and a particular phoneme can be described using a vector of PFs. By decomposing phonemes into a universal set of PFs, we can avoid issues of mismatched phoneme inventories across multiple languages and more efficiently share acoustic information between them when pooling training data.

In this paper, we present in detail the TTS system we submitted for the Blizzard Challenge 2021 spoke task. Our model is based on the FastPitch architecture [6] with additional word-level language embeddings, trained with multiple speakers of Spanish and English using phonological feature inputs. At synthesis time, words automatically identified as English are nativized to Spanish phonology through a set of hand-written rules. In addition to reporting Blizzard test results, we perform a supplementary evaluation where we focus on comparing the effects of using phoneme symbols vs. phonological features and different nativization strategies, while using ground-truth language identification labels.

## 2. Data

We trained our system using data from four different speech corpora, summarized in Table 1.

### 2.1. Blizzard Challenge 2021

The Blizzard Challenge 2021 data set consists of one Peninsular Spanish female speaker with 9.5 hours of audio data and matching transcriptions, manually checked. We trimmed silences at the start and end to normalize them all to 500 milliseconds, which was necessary to obtain correct forced alignments of the data, and reduced the total audio duration to 6.4 hours. This data set contains exclusively Spanish utterances.

### 2.2. L2-ARCTIC

In order to obtain speech samples of English with a Spanish accent, we made use of the L2-ARCTIC dataset (v5.0) [7]. This corpus includes recordings from 24 non-native speakers of English with different first-language backgrounds including four Spanish speakers, two male and two female. We used only the two female Spanish L1 speakers *MBMPS* and *NJS*, to match the gender of the Blizzard speaker. Although the amount of data per speaker is relatively low (about an hour each), because the speakers are reading the ARCTIC script [8], the available data has a balanced phonemic coverage of English. Although the corpus includes other materials such as manual annotations of pronunciation errors, we did not make use of them.

Table 1: *Summary of training data used, including number of utterances and total duration in hours.*

| Corpus | Language | Speakers | Utts | Hours |
|---|---|---|---|---|
| Blizzard 2021 | Spanish | 1 | 4900 | 6.4 |
| L2-ARCTIC | English | 2 | 2243 | 2.3 |
| LJ Speech | English | 1 | 2250 | 4.1 |
| Common Voice | Spanish | 62 | 5226 | 7.3 |
|  | Total | 66 | 14619 | 20.0 |

### 2.3. LJ Speech

We randomly sampled a subset of the LJ Speech corpus [9] to match the number of utterances in the L2-ARCTIC data. This data provides examples of native American English pronunciation. We decided to include American English data (rather than e.g. British English) based on some pronunciations of English words by the target Blizzard speaker in the small development set provided, for example with post-vocalic /ɹ/ being pronounced in the word ‘computer’.

### 2.4. Mozilla Common Voice

The Mozilla Common Voice (MCV) project collects crowd-sourced speech recordings with associated text transcriptions across multiple languages [10]. As a language with a large web presence, Spanish text prompts for volunteer recording are largely extracted from Wikipedia articles. We expected this corpus might provide a good source of Spanish utterances with some degree of code-switching, for example with some proper nouns included from English.

We selected a subset of validated utterances from the v6.1/2021-12-11 release of the Spanish portion of Common Voice. Volunteers may provide some demographic information, including self-reported accent and gender labels. We sampled between 50–100 utterances from speakers with at least 50 utterances and Peninsular accent labels. We excluded from consideration any speakers where manual review of 5 random utterances found multiple different speakers associated with the same ID, excessive background noise, poor microphone quality or where recordings appeared to have been made at a lower sampling frequency than 22.05 kHz (based on inspection of spectrograms). These criteria allowed us to approximately match the number of utterances in the single-speaker Blizzard data set, with 5226 utterances sampled from 12 female and 50 male speakers. We found that introducing a large number of additional Spanish speakers improved the ability of our model to maintain the Blizzard speaker’s voice quality when switching language embeddings between Spanish and English words in the test utterances.

## 3. Data processing

### 3.1. Forced alignment and phonemic transcription

FastPitch requires explicit duration information during training, for which we used forced alignments generated using the Montreal Forced Aligner (MFA) [11]. We trained a new acoustic model for each corpus included in our training data (including speaker adaptation for multi-speaker corpora) rather than using pre-trained Spanish or English models provided with MFA.

To align the Blizzard data, we used the available MFA Spanish grapheme-to-phoneme (g2p) conversion model to obtain phonemic transcriptions in the GlobalPhone phone set [12].

Table 2: *List of distinctive phonological features.*

| Category | Features |
|---|---|
| Major class | syllabic, consonantal, sonorant |
| Cavity | coronal, anterior, distributed, labial, high, low, back, round, nasal, lateral, constricted glottis |
| Manner | continuant, delayed release, tense, long |
| Source | voice, strident, subglottal pressure |
| Other | silence |

We made a small correction when we noticed that many words ending in <-ado> or <-ido> (suffix for participles) were transcribed with the intermediate consonant deleted. Although this is a valid pronunciation for some Spanish speakers, it did not reflect the pronunciation of the Blizzard speaker and as such the intermediate /d/ was automatically inserted for all corresponding words.

For the English L2-ARCTIC and LJ-Speech corpora, phonemic transcriptions were generated either by lookup in the American English (GAM) version of the Combilex lexicon [13] or using a decision tree-based g2p implemented for the same phone set in the Festival TTS system [14].

For the Spanish Mozilla Common Voice corpus, alignments were again made against phonemic transcriptions generated from the MFA g2p model. Checking a sample of 500 utterances suggested that around 10% might contain English words, representing around 2% of individual word tokens. However, we were not able to account for this early in the data preparation, so that automatically generated Spanish pronunciations were used also for any English words in the corpus, potentially impacting the quality of the resulting alignments.

### 3.2. Phonological features

After converting text to phonemic representations, we expanded atomic phoneme symbols into phonological feature vectors. We first mapped GlobalPhone and Combilex phones to their corresponding symbols in the International Phonetic Alphabet (IPA) [15], then to binary PFs using PanPhon [16]. This library converts potentially very fine-grained phonetic transcriptions including IPA diacritics to a set of 21 articulatory features similar to those laid out in [5]. During forced alignment, we split any diphthongs into sequences of two vowel symbols to enable this conversion; affricates are still handled as single symbols by PanPhon. To this set we added a single additional feature to represent any silences inserted by MFA, for example introduced by punctuation. Table 2 lists the feature set used; values are $\pm 1$ or 0 where some features are unspecified for certain segments, e.g. vowel height features on consonants.

We noted two deficiencies in the set of phonological features used. First, the featural representations of the alveolar trill /r/ provided by PanPhon is identical to that of the alveolar tap /ɾ/, masking a phonemic contrast in Spanish. This caused our submitted system to make some segmental mistakes, sometimes producing /ɾ/ (the more common sound in our training data) where /r/ would have been the correct phoneme. Second, there are no features accounting for lexical stress. We merged symbols for stressed and unstressed vowels before conversion to PF vectors, also resulting in some stress assignment errors in our submitted samples.

### 3.3. Language identification

We generated word-level language labels for Blizzard evaluation test utterances using the following procedure. First, we pass the whole utterance through a Spanish-English language identification model [17]. This is a pre-trained model that uses a multilingual BERT trained with the LinCE corpus [18], a benchmark for code-switching tasks. We did not fine-tune the model to our data. The model outputs a label for every token in the input string, with four possible values: Spanish, English, named-entity or other (for punctuation). Many of the words of English origin in the test data were identified as named-entities, so these outputs required further disambiguation.

Second, we checked every word labelled as a named-entity for characters exclusive to Spanish: á, é, ú, í, ó and ñ. If any of these were present, the word was classified as Spanish. Otherwise, the word was passed to a second pre-trained model for multiple language detection based on N-gram probabilities [19]. If the word was identified as belonging to a Romance language, then it was classified as Spanish, otherwise as English.

To calculate the accuracy of the labels, we randomly selected a set of 40 utterances from the test data, ensuring that every sentence had at least one word in English. Of all words, 94.7% were correctly classified, although only 17.5% came from English. Of English words, 73.2% were correctly identified, compared to 99.3% of the Spanish words.

### 3.4. Pronunciation nativization rules

For test utterances, we converted input text to phonemes using our own rule-based g2p system for detected Spanish words. For English (or non-Spanish) words, we first retrieved English pronunciations using Combilex resources as for forced alignment. Then, we applied a series of nativization rules to bring English pronunciations more in line with Spanish phonology, for example inserting an epenthetic /e/ before word-initial /s/, converting English-only fricative /ʃ/ to shared /tʃ/ and converting /ə/ to full Spanish vowels based on orthography, e.g. 'Twitter' /twɪtəɹ/ → /twiteɹ/. Note that we maintained the English approximant rhotic /ɹ/ rather than mapping to Spanish tap or trill. We also preserved many word-final consonant clusters which typically do not occur in native Spanish words.

While we tried to produce a consistent and comprehensive set of English-Spanish transformation rules, there are inevitably exceptions which cannot be handled. This may also be exacerbated by errors in the initial English g2p, itself a challenging task, feeding into our transformation rules. We did not manually fix any mistakes at the end of the full g2p pipeline for the submitted test stimuli, since this would not be possible in a production scenario.

## 4. System description

Our base model architecture is FastPitch [6], building on the reference implementation from NVIDIA. FastPitch is a sequence-to-sequence transformer based model, with explicit duration and fundamental frequency prediction. Symbol embeddings are summed with positional embeddings as input for the transformer encoder. For PF inputs, we replace the phoneme embedding table with a linear layer projecting to hidden representations of matching dimension (512). Because we are training our models with multiple speakers, an additional speaker embedding is added to this sum. Finally, a language embedding is also summed, where the Blizzard data and MCV datasets are encoded as Spanish and LJ Speech and L2-ARCTIC as English.

Encoder outputs are used to predict normalized fundamental frequency and duration in frames per symbol. At training time, ground-truth values are provided by fundamental frequency contours extracted with Parselmouth [20] and durations from MFA, as described in Section 3.1. The predicted fundamental frequency is embedded and summed with the encoder outputs. The result is upsampled using predicted durations and input to the transformer decoder. The decoder output is projected to the mel spectrogram dimensionality.

Our final submitted system used PF inputs derived from phoneme strings generated using our Spanish g2p and nativization rules for English words. Language embeddings were specified word-by-word according to our automatic language ID system. The model was trained for 1000 epochs on the combined data from all corpora as listed in Table 1. To convert generated mel spectrograms back to the waveform domain we used Waveglow [21], with a pre-trained model from NVIDIA fine-tuned to the target Blizzard speaker for 450 epochs. Audio samples are available online.[1]

## 5. Blizzard Challenge 2021 evaluation

Systems submitted to the spoke task of the Blizzard Challenge 2021 were evaluated by subjective listening test along three dimensions: overall naturalness of synthesized speech, similarity to the original Blizzard data speaker and acceptability of the English words included in mostly Spanish utterances. Figures 1 and 2 present box plots of system ratings for naturalness and acceptability of English words respectively (we omit speaker similarity results for brevity, since we do not focus on this aspect in our own evaluation). Our system (H) is outlined in each plot; system R represents natural speech from the Blizzard speaker. Each plot includes responses from all listeners engaged in the Blizzard evaluation. Differences between systems were tested for significance using Wilcoxon's signed rank tests (see [22] for more details).

Our system received a mean rating of 2.86 for overall naturalness (not significantly different from systems M and D), 2.8 for speaker similarity (not significantly different from A, C, D and M) and 2.6 for acceptability of English words (not significantly different from N).

## 6. Supplementary listening test

We conducted an additional listening test explicitly comparing our phoneme- and PF-based systems, using different nativization strategies. To avoid the influence of any language ID errors, we used the same sample of 40 utterances from the Blizzard test set used to evaluate our language ID system in Section 3.3. We synthesized the test utterances using the following nativization strategies for each of phoneme and PF inputs (240 stimuli total):

- *Mixed (mix)* – Pronunciations and word-level language embeddings determined by ground-truth language labels
- *Nativized (nat)* – Pronunciations determined by ground-truth language labels, all language embeddings set to Spanish
- *Spanish (esp)* – All pronunciations produced by Spanish g2p rules only (no nativization of English words) and all language embeddings set to Spanish

We expected these systems to range from the most faithful renditions of English words in *mix*, through more heavily Spanish-

[1] `https://dan-wells.github.io/blizzard2021`

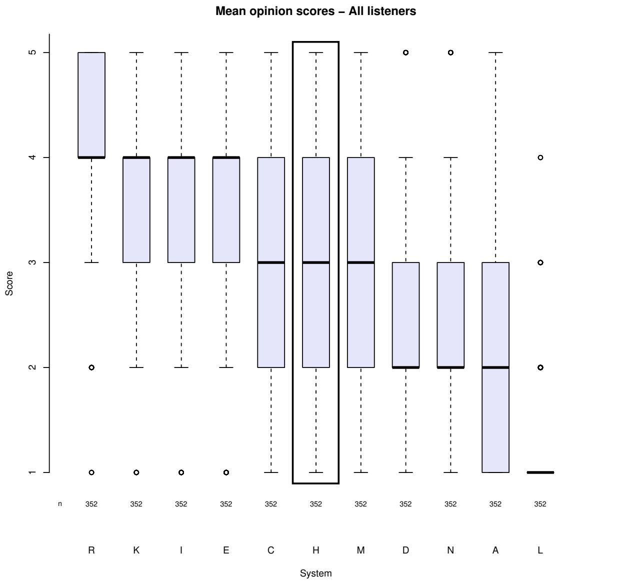

Figure 1: *Mean opinion scores for all systems submitted to Blizzard evaluation (ours system H).*

accented in *nat* and finally to English interpreted strictly according to Spanish orthographical rules in *esp*. Note that the *mix* condition with PF inputs is equivalent to an oracle language ID version of our submitted system.

We recruited 40 native speakers of Spanish (located in Spain) through Prolific. Each participant listened to a random sample of 40 synthesized utterances from the 6 systems under test. Each synthesized audio file was presented individually along with its transcription, with English words highlighted using bold text. We asked participants to respond to three statements on a 5-point Likert scale from "Disagree" to "Agree":

1. This phrase is very natural.
2. Considering the marked words, this speaker shows mastery of English.
3. I would pronounce the English words the same way.

The average test duration was 19 minutes, and participants were paid £5 for their time. In total, we gathered 1600 ratings for each question, with each stimulus rated by 6 or 7 participants and each system rated an average of 266 times.

### 6.1. Results

Figures 3–5 show box plots for responses to questions 1–3 respectively. Labels for systems using phonological feature inputs are prefixed with *F* and phonemes with *P*. Systems are ordered by mean naturalness rating, but following [22] we make no comparisons on that basis.

We tested for significant differences between systems using pairwise two-sided Mann-Whitney U tests, judged at the 1% level (Bonferroni-corrected $p < 0.00067$ for 15 comparisons). Considering naturalness ratings for pairs of systems using the same nativization strategy but with different input representations, there were no significant differences: PF and phoneme inputs received similar scores. Both *F-Mix* and *F-Nat* produced significantly more natural speech than *F-Esp*, but were not significantly different from each other. There were no significant differences between any of the phoneme-based systems.

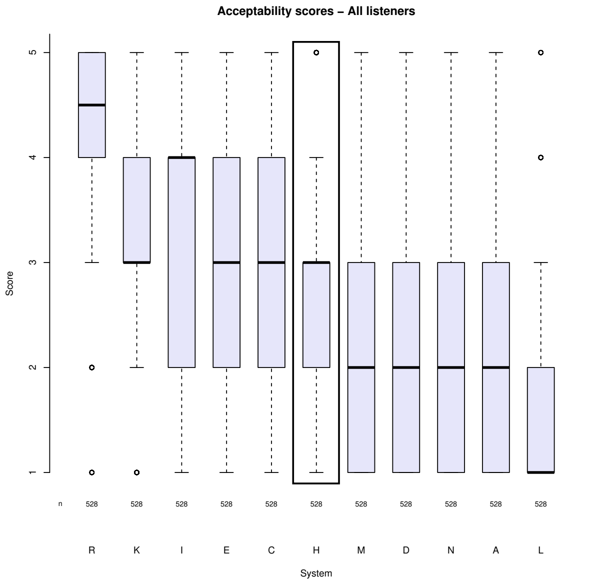

Figure 2: *Acceptability of English words in test utterances for all systems submitted to Blizzard evaluation (ours system H).*

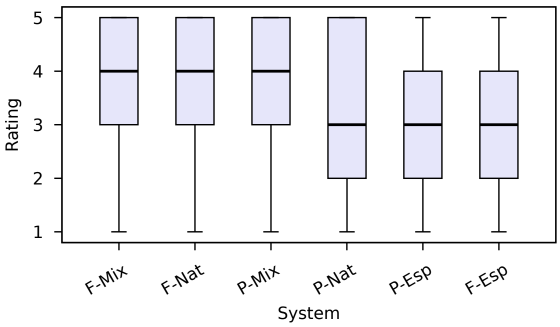

Figure 3: *Naturalness ratings for systems with different input representations and nativization strategies.*

We consider the lack of difference between input representations to be a result of the application of nativization rules. Assimilating test utterances to Spanish phonology also removed most of the English-only phonemes which might be expected to benefit most from the parameter sharing enabled by our phonological feature representations (for example any vowels which occur only in the smaller English portion of our training data). The only remaining English phonemes were /ɹ/ and /ə/, representing 1% and 0.3% of all phonemes in the test utterances respectively. During system development we found the introduction of nativization rules to help considerably compared to passing full English phone sequences through even a PF-based system, perhaps due to lack of training data with English phones surrounded by Spanish context, so we considered this trade-off to be worthwhile.

For ratings of the synthetic speaker's perceived mastery of English and similarity of the pronunciations of English words compared to listening test participants' own, there were again no significant differences between systems using the same nativization strategy but different input representations. Within

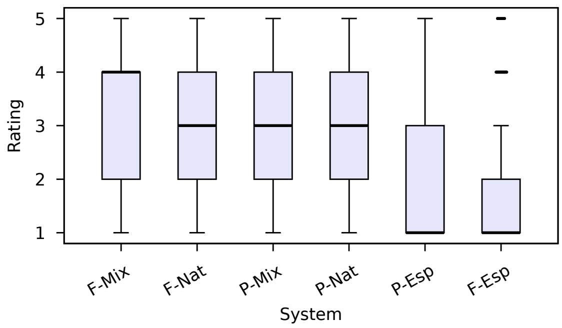

Figure 4: *Ratings of speaker's knowledge of English for systems with different input representations and nativization strategies.*

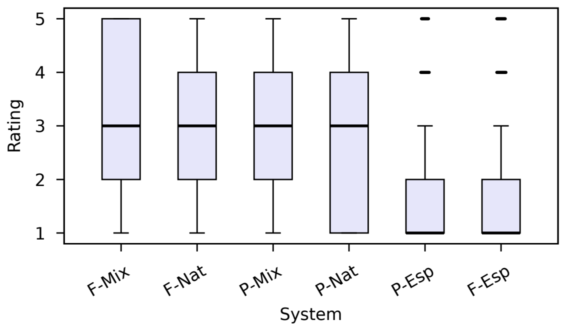

Figure 5: *Similarity ratings for English words compared to listener's pronunciation for systems with different input representations and nativization strategies.*

both phoneme- and PF-based systems, the *mix* and *nat* strategies were rated significantly higher than *esp*, but again were not judged significantly differently from each other.

The overall relationship between naturalness ratings and perceived knowledge of English or listener pronunciation similarity varied with the nativization strategy but was not very strong in any case. We calculated Spearman's rank correlation coefficients ($\rho$) between responses to our naturalness question and the two English-focused questions separately, giving values between 0.40 and 0.53 for both PF- and phoneme-based *mix* and *nat* systems and between 0.18 and 0.24 for *esp* systems. On the other hand, correlations between apparent English knowledge and pronunciation similarity were quite high, $0.71 \leq \rho \leq 0.85$ for all systems. This suggests that listeners were evaluating the whole sentence for naturalness, including Spanish words and prosody. For example, Blizzard test utterance 0026 synthesized by system *F-Nat* received an average naturalness rating of 2.5, while averaging 4.0 on the other two questions. After listening to this sample, we hypothesize that this is due to errors in the Spanish part of the sentence: the presence of the mispronunciation of /r/ at the start of a word and unnatural intonation.

The two questions regarding the English words were designed to see if the listeners could decouple their own pronunciations from what they consider to be a knowledgeable speaker of English. Overall, listeners preferred to align their pronunciation with the strategy that was closer to English, i.e. *mixed*. However, for some cases we saw a mismatch between their own pronunciations and what was considered a knowledgeable pronunciation. For example, utterance 0076 from system *P-Esp* had an average rating of 3.0 on knowledge of English and 3.8 on pronunciation similarity to the listener. Interestingly, the English word in this utterance was "spoiler", one of the most frequent English words used in Spanish identified by [2]. In this system, even English words are synthesized with a Spanish language embedding, and here "spoiler" is pronounced with a very strong /e/ sound before the initial /s/, complying with Spanish syllabic structure (and a very common pronunciation among native speakers of Spanish). This utterance was rated 2.4 for naturalness on average; the Spanish words had a strange rhythm and intonation. Utterance 0194 from system *P-Mix* showed the inverse pattern, with an average rating of 4.3 for English knowledge but 3.1 for pronunciation similarity (with a naturalness of 3.1). The English words in this utterance were "circuit breaker" (not a very common term in everyday Spanish), with a pronunciation close to the English especially for the vowels, where a Spanish speaker would more commonly map them to near Spanish equivalents.

# 7. Conclusion

We used phonological feature vector representations to streamline multi-lingual TTS training for synthesizing Spanish utterances also containing some English words in our submission to the spoke task of the Blizzard Challenge 2021. Our final system also incorporated a series of nativization rules informed by automatic language identification over test sentences, through which we brought the pronunciations of detected English words more in line with Spanish phonology. While application of these rules somewhat negates the potential benefits of using PFs during model training by effectively collapsing test utterances to the Spanish phoneme inventory, we nonetheless found it beneficial for our submitted system. Given that our systems were trained on data combining native and non-native speakers of English and native speakers of Spanish, but without any explicit code-switching labels, we found our performance in the Blizzard Challenge evaluation encouraging. In future work, we want to systematically test which of these data sets brought the greatest improvement to our systems.

We also conducted a supplementary listening test which showed that evaluating preferences between nativization strategies is a complicated task. Still, we can draw some conclusions from our analysis of results from this test. Our evaluation is limited by the quality of the Spanish portions of utterances synthesized by our system: to evaluate properly only the nativization of English words, we would first need perfect generation of Spanish. The case analysis shows that although listeners tend to report that their own pronunciations align with systems that show a pronunciation closest to English, they can still notice in some cases that these don't match. If we want to obtain a transparent preference from listeners with respect to nativization strategy, we need to take this into account with careful design of test sentences, considering especially the frequency of use of the target English words by listeners in their own everyday speech.

**Acknowledgements:** Dan Wells is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. Pilar Oplustil-Gallegos is supported by: ANID, Becas Chile, nº 72190135.

# 8. References

[1] T. Polyakova and A. Bonafonte Cávez, “Nativization of english words in spanish using analogy,” in *Proceedings of the 7th ISCA Speech Synthesis Workshop*, 2010, pp. 294–299.

[2] R. Janssen, “La influencia del idioma inglés en el idioma español y la medida en que los españoles utilizan los anglicismos en el lenguaje cotidiano,” B.S. thesis, Utrecht University, 2020.

[3] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, “Phonological Features for 0-Shot Multilingual Speech Synthesis,” in *Interspeech 2020*. ISCA, 2020, pp. 2942–2946.

[4] A. Gutkin, “Uniform Multilingual Multi-Speaker Acoustic Model for Statistical Parametric Speech Synthesis of Low-Resourced Languages,” in *Interspeech 2017*. ISCA, 2017, pp. 2183–2187.

[5] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper & Row, 1968.

[6] A. Łańcucki, “Fastpitch: Parallel Text-to-Speech with Pitch Prediction,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[7] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-arctic: A non-native english speech corpus,” in *Proc. Interspeech*, 2018, p. 2783–2787. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1110

[8] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *Fifth ISCA workshop on speech synthesis*, 2004.

[9] K. Ito and L. Johnson, “The LJ Speech Dataset,” 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset

[10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common Voice: A Massively-Multilingual Speech Corpus,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4218–4222.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner (version 2.0.0 alpha 7),” 2021. [Online]. Available: https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[12] T. Schultz and T. Schlippe, “GlobalPhone: Pronunciation Dictionaries in 20 Languages,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. European Language Resources Association (ELRA), 2014, pp. 337–341.

[13] K. Richmond, R. A. J. Clark, and S. Fitt, “Robust LTS Rules with the Combilex Speech Technology Lexicon,” in *Interspeech 2009*, 2009, pp. 1295–1298.

[14] R. A. J. Clark, K. Richmond, and S. King, “Festival 2 – Build Your Own General Purpose Unit Selection Speech Synthesiser,” in *5th ISCA Speech Synthesis Workshop*, 2004, pp. 173–178.

[15] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[16] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, “PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016, pp. 3475–3484.

[17] S. Sarker, “CodeSwitch,” 2020. [Online]. Available: https://github.com/sagorbrur/codeswitch

[18] G. Aguilar, S. Kar, and T. Solorio, “LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation,” in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 1803–1813.

[19] “PyLaDe - Language Detection tool.” [Online]. Available: https://github.com/fievelk/pylade

[20] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A Flow-based Generative Network for Speech Synthesis,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[22] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the Blizzard Challenge 2007 listening test results,” in *The Blizzard Challenge 2007*, 2007, pp. 1–6.