Cross-Lingual Neural Network Speech Synthesis Based on Multiple Embeddings

Tijana V. Nosek[1]*, Siniša B. Suzić[1]*, Darko J. Pekar[2], Radovan J. Obradović[2], Milan S. Sečujski[1], Vlado D. Delić[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad (Serbia)
[2] AlfaNum – Speech Technologies Ltd., Novi Sad (Serbia)

Received 17 December 2020 | Accepted 14 June 2021 | Published 12 November 2021

Abstract

The paper presents a novel architecture and method for speech synthesis in multiple languages, in voices of multiple speakers and in multiple speaking styles, even in cases when speech from a particular speaker in the target language was not present in the training data. The method is based on the application of neural network embedding to combinations of speaker and style IDs, but also to phones in particular phonetic contexts, without any prior linguistic knowledge on their phonetic properties. This enables the network not only to efficiently capture similarities and differences between speakers and speaking styles, but to establish appropriate relationships between phones belonging to different languages, and ultimately to produce synthetic speech in the voice of a certain speaker in a language that he/she has never spoken. The validity of the proposed approach has been confirmed through experiments with models trained on speech corpora of American English and Mexican Spanish. It has also been shown that the proposed approach supports the use of neural vocoders, i.e. that they are able to produce synthesized speech of good quality even in languages that they were not trained on.

Keywords

Cross-lingual, Neural Networks, Speech Synthesis, Vocoder.

DOI: 10.9781/ijimai.2021.11.005

I. Introduction

MODERN text-to-speech (TTS) systems should not only be able to produce intelligible and natural-sounding speech but also to produce speech in multiple voices, styles, and preferably in multiple languages as well. Any TTS system which can handle input text in more than one language, and produce speech based on it, is referred to as multilingual TTS [1]. Multilingual TTS systems have a wide range of application. Besides being used within speech-to-speech language translation systems and interactive language tutoring systems, this functionality is necessary for a TTS system to be able to insert an occasional foreign language word into otherwise mono¬lingual speech, or even to alternate between languages in a manner consistent with the syntax and phonology of each language, which is referred to as code mixing.

The simplest solution to multilingual synthesis is based on simultaneous use of separate monolingual systems. However, since such systems are typically trained on speech corpora recorded by different speakers, this inevitably leads to inferior quality in code-mixing scenarios. On the other hand, TTS systems which can produce multiple languages in the voice of a single speaker, but typically require speech corpora from bilingual or polyglot speakers, are referred to as polyglot TTS [1], [2]. The capability to produce speech in a particular language although the training speech data in the voice of the target speaker does not contain any speech in that language is referred to as cross-lingual speech synthesis. It represents a natural alternative to using polyglot speakers for the production of training data, brought about by scientific and technological development in machine learning.

Initial attempts at producing cross-lingual synthesis were proposed for concatenative systems in the early 2000s and were based on creating phoneme mappings between source and target language [3]–[5]. Such approaches were able to generate phonetically accurate speech output, but since the intonation they were able to achieve was based on the existing source language, they were mostly applicable in code-mixing scenarios to generate some foreign words. Furthermore, sufficiently accurate phoneme mappings could be established only between languages with similar phonetic content. Another approach to concatenative cross-lingual synthesis is based on frame-level mapping. In the algorithm proposed in [6], source speaker recordings in language L1 are first spectrally warped towards target speaker recordings in language L2. Warped trajectories from source speaker are used for guiding the selection of appropriate frame-level spectral features from the target speaker, resulting in a set of utterances in L1 made from frames belonging to the target speaker. Selected target speaker frame features are then used for training a Hidden Markov Model (HMM) system capable of producing speech in the voice of the target speaker, but in L1, a language absent from the initial speech corpus of the target speaker. An extension of this approach is described in [7], proposing the use of bilinear spectral warping instead of piecewise linear, inclusion of original speech from target speaker

* Corresponding author.
E-mail addresses: tijana.nosek@uns.ac.rs (T. V. Nosek), sinisa.suzic@uns.ac.rs (Siniša B. Suzić).

in L2 into the training set for the HMM, as well as joint treatment of phonemes from both L1 and L2 based on their places and manners of articulation.

The shift of the focus in speech synthesis from concatenative to parametric approaches, brought about by the need for increased flexibility, has also influenced the development of cross-lingual speech synthesis. The first such approach, based on HMM, has enabled cross-lingual synthesis based on state mapping [8]. In this approach a bilingual speaker corpus is used to create two decision trees and an appropriate mapping between their terminal nodes is then established based on KL divergence. The obtained mapping is then applied to a monolingual speaker to generate speech in a new language. A language conversion method based on a mapping between terminal nodes of two decision trees created for average voice models is presented in [9]. A framework which attempts to factorize speaker and language features, which are modelled using a range of transforms, is presented in [10].

A major breakthrough in the development of high-quality parametric TTS did not come until the advent of neural networks. Scientific progress in this area has led to a number of different approaches to cross-lingual speech synthesis as well. For instance, in the research proposed in [11] acoustic features used to produce speech in the target language are created by a deep bidirectional long short-term memory (DB LSTM) network on the basis of phonetic posteriorgrams (PPG). The network is trained using original acoustic features of the target speaker as well as PPGs of the target speaker in the source language, obtained by a speaker-independent automatic speech recognition (ASR) system in the target language. Synthesis involves input of an arbitrary text to a general TTS in the target language (trained on any non-target speaker), which is then converted into a PPG by the ASR. The PPG features are then fed to the DB LSTM, which generates acoustic features of the target speaker in the target language, according to the input text. In [12] a deep neural network (DNN) based ASR is used to match senones from one speaker-dependent HMM TTS in the source language and another one in the target language. An example of multi-speaker and multi-language DNN TTS model is presented in [13]. This model uses separate input layers for each language and separate output layers for each speaker, while hidden layers are shared among all speakers and languages, and cross-lingual synthesis is achieved by combining corresponding input language layers and output speaker layers. In [14] unsupervised adaptation of multilingual TTS is performed by way of a search for a linguistic context which matches the available acoustic features to the greatest degree. It has been shown that a multi-speaker architecture in language L2 can be adapted by using speech data from a single speaker in language L1 to obtain TTS in language L2 in the voice of the target speaker.

Recently, end-to-end systems, which enable speech to be produced directly from text, have achieved remarkable results [15]–[17], but they require very large quantities of training data to produce synthetic speech of high quality. The end-to-end approach has also been introduced into the area of cross-lingual TTS. Most notable approaches to cross-lingual end-to-end speech synthesis based on Tacotron2 were presented in [18]–[20]. In [18] and [19] the Tacotron2 model is extended with speaker, language, tone and stress embeddings, while [18] introduces an additional adversarial speaker classifier and residual encoder. A speaker encoder based on ResCNN architecture, used for creating embeddings which condition the Tacotron2 system for predicting spectral envelopes, is presented in [20]. In all three methods shared IPA representations of phonemes are used. In [21] speaker embeddings for bilingual speakers are analysed and it has been shown that these embeddings form distinct, partly overlapping clusters. Cross-lingual speech synthesis is obtained by applying a translation of cluster embeddings learned from a bilingual speaker to a monolingual one using a Tacotron based architecture.

In spite of their great potential, a major drawback of end-to-end systems is their requirement not only for extreme computational power but also for very large quantities of speech data, which is a problem for under-resourced languages. The model that will be presented in the paper enables high quality speech synthesis, even in cross-lingual scenario, with very limited resources. It is evaluated through 5 listening tests, examining (1) whether the quality of synthesis decreases in comparison with monolingual TTS; (2) how the quality of synthesis in the original language by the proposed model compares with cross-lingual synthesis; (3) whether voice characteristics remain preserved in the cross-lingual scenario; (4) to what extent synthesis quality is degraded when the multilanguage model is adapted to a new speaker; and (5) how a neural vocoder compares to a deterministic one in the cross-lingual scenario.

The remainder of the paper is organized as follows. In Section II, we present a novel method and architecture for speech synthesis in multiple languages, in voices of multiple speakers and in multiple speaking styles, as well as speech data used in the training and evaluation of the proposed method. In Section III we give a detailed presentation of the experiments and their results. In Section IV we discuss the results of the experiments and in Section V we draw appropriate conclusions about the performance of the proposed method and outline the directions of future work.

II. Methods

This section will give a detailed presentation of the architectures of the models used in the experiments, training data, as well as specific points related to the implementation of models.

A. Models

The model proposed in this research builds upon our previous solution for monolingual speaker/style dependent speech synthesis based on embedding [22]. Both models follow the standard structure of speaker-dependent TTS, which will be outlined below.

1. Standard Speaker-Dependent TTS

A standard speaker-dependent TTS system consists of two neural networks, one for predicting phonetic segment durations, and the other for predicting acoustic features for each frame. The inputs of both networks contain linguistic information extracted from phonetically and prosodically annotated text. In order to take into account phonetic context, the inputs of both networks include not only the phonemic identity of the current phone, but the identities of phones at positions from −2 to 2 relative to the current phone. Phonemic identities are presented to the network as one-hot vectors, with some obvious exceptions, e.g. if a phone is sentence initial, features related to positions −2 and −1 are undefined and hence represented by all-zero vectors. Individual prosodic features are also presented to the network as additional inputs in binary form, and each of them represents an answer to a yes/no question typically related to the type and position of a particular prosodic event with respect to the current phone (such as: “Is the current phoneme in a stressed syllable?”, “Is the number of syllables until the next phrase break greater than 3?” etc.). The acoustic network also obtains the information regarding phone durations and position of the current frame relative to its HMM state. In the synthesis phase, the outputs from the duration network are used as additional inputs to the acoustic network. The outputs of the acoustic network are typically converted to synthetic speech waveforms using an appropriate vocoder. A number of approaches have been proposed to extend such a model to enable it to handle multiple speakers and/or speech styles [22]–[24], or to adapt it to a certain speaker and/or speech style [14], [22], [25].

2. Monolingual Speaker/Style-Dependent TTS Based on Embedding

The model used as a starting point in this research, represents an extension of the standard speaker-dependent TTS, which supports multiple speakers and styles and requires a very limited quantity of speech data for adaptation [22]. This model follows the basic structure of the standard speaker-dependent TTS in that it is based on two neural networks, one predicting phone durations and the other predicting frame-level acoustic features, both using phonetic transcriptions and prosodic features as inputs.

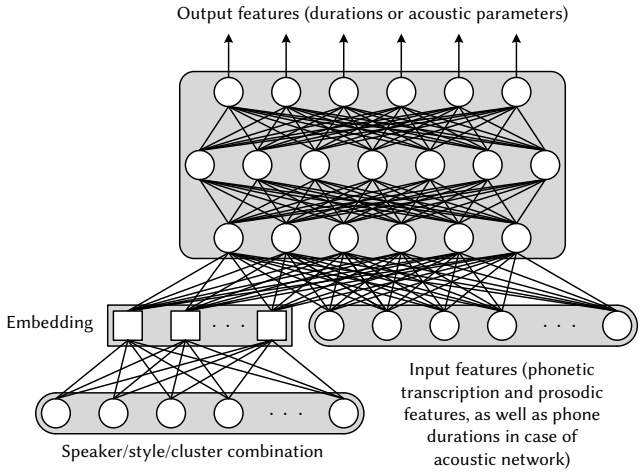


Fig. 1. Architecture of either of the two neural networks that predict phone durations or acoustic features in monolingual speaker/style-dependent TTS based on embedding (MTTSE) [22].

The model presented in [22], capable of adaptation to different speakers and styles, uses an embedding vector as supplementary information at the input of both acoustic and duration networks, as shown in Fig. 1. This embedding vector, obtained in the training phase, uniquely represents a particular combination of speaker, speech style and cluster (a portion of the training data consistent in acoustic and prosodic quality, which typically corresponds to a single recording session). In other words, speaker ID, style ID and cluster ID are jointly represented as a single one-hot vector, which is converted into an appropriate joint embedding through training. In the resulting low-dimensional embedding space created by the network, the distance between points representing particular speaker/style/cluster (SSC) combinations reflects their similarity in terms of acoustic features and speech rate, which helps the network to efficiently generalize on unseen speech data. Hereafter, this model will be referred to as "monolingual text-to-speech based on embedding" (MTTSE).

3. The Proposed Model

The essential problem in a multilanguage scenario arises from the discrepancies between linguistic features across languages. To begin with, two languages generally do not share the same phonological inventory. Although it is usually possible to identify certain phonemes as common to multiple languages in a cross-lingual scenario, treating them as such can have negative effects since there may still be slight differences at the phonetic level. For that reason, the proposed model treats all phonemes from all languages as separate entities, which are uniquely represented as one-hot vectors, and then embedded into a low-dimensional space, as was the case with unique SSC combinations in MTTSE. The idea behind this approach is that the distance between points in the phonetic embedding space should reflect the degree of similarity between corresponding phones regardless of their language. The proposed model, hereafter referred to as "cross-lingual text-to-speech based on embedding" (CTTSE), uses 5 different embeddings for each phone in the corpus, and they are related to the phonemic identity of phones at positions from −2 to 2 relative to the current phone, as shown in Fig. 2. The size of this vector equals the sum of the sizes of phoneme inventories of all languages covered by the system, and phonetic embedding achieves efficient dimensionality reduction. Such an approach allows the network to decide e.g. to what degree the English /s/ and the Spanish /s/ are similar, and no expert knowledge is needed to match phonemes across languages.

As is the case with MTTSE, besides phonetic features, both the duration network and the acoustic network require prosodic features at their inputs. The proposed model assumes that the same prosodic annotation scheme is used for all languages included in the training. For that reason, it was possible to consider a great majority of prosodic features to be common between languages and to present them to the

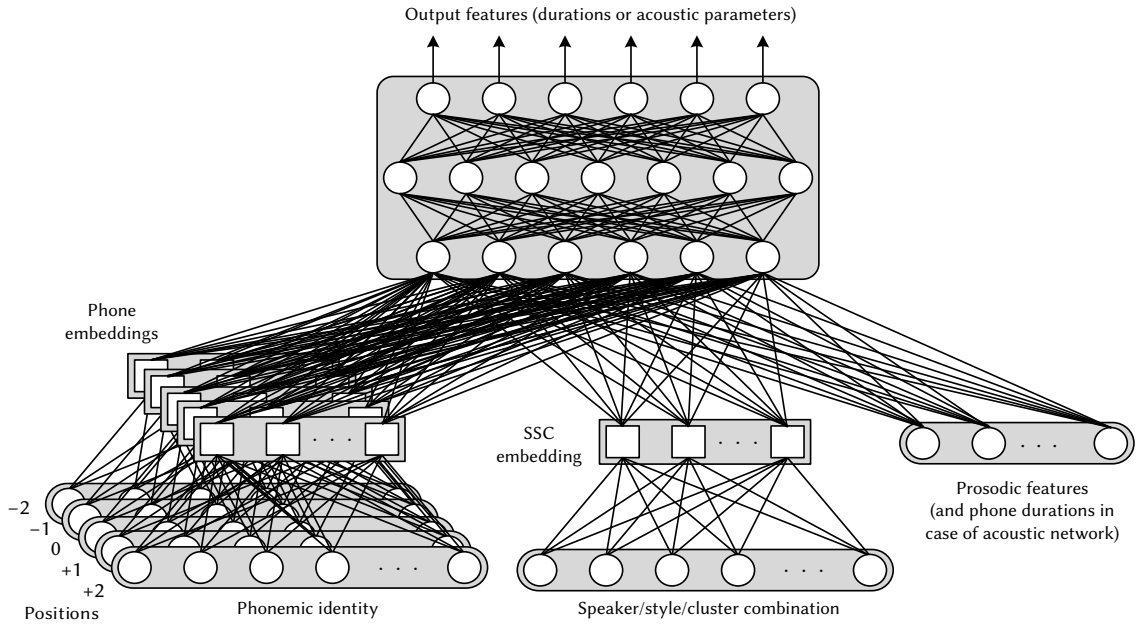


Fig. 2. Architecture of either of the two neural networks that predict phonetic segment durations or acoustic features within the model proposed for cross-lingual speaker/style-dependent TTS based on embedding (CTTSE).

neural network directly, as shown in Fig. 2. If this were not the case, prosodic inventories would also be different across languages, and some form of prosodic embedding would have been necessary.

The proposed model not only allows multi-lingual speech synthesis but cross-lingual synthesis as well. Namely, using a specific speaker-style embedding at synthesis stage will produce the voice of the desired speaker in the desired speaking style regardless of the language of the input text. Since the information of phonetic features, prosodic features and voice-related features are separated, they can be used in different combinations, enabling synthesis in the voice of a speaker who has never been "heard" speaking the target language, i.e. whose training data do not contain any speech in the target language.

Besides being able to perform cross-lingual speech synthesis, the proposed model is also capable of adaptation to the voice of a speaker not seen in the training phase, just as was the case with MTTSE, based on adaptation speech data in any of the languages in the initial model. The first phase is aimed at establishing the embedding for the new speaker/style, and it begins by random initialization of the values in the embedding layers of both networks. In this phase of the adaptation process, only the values in the embedding layers of both networks are adjusted through back-propagation while the rest of the networks is kept unchanged. The model with embedding layers adapted in such a way synthesizes speech that already resembles the target speaker/style to some extent. However, the quality of synthesized speech can be further improved through the second phase of the adaptation process, in which the same training data is used again, but the embedding layer is frozen, while the weights in the networks are modified according to the back-propagated error.

B. Data

The speech data used in this research include two languages, American English and Mexican Spanish, with a total duration of about 31 hours. All recordings were sampled at a rate of 22.05 kHz and 16 bits per sample were used. Some recordings were made in professional studios, while others were obtained from publicly available audiobooks or speeches. In the data of each speaker, styles and clusters are identified. The entire speech corpus is phonetically and prosodically annotated, and the prosodic annotation of both languages follows the Tone and Break Indices (ToBI) set of conventions, with certain extensions related to the degree of acoustic realization of pitch accents, as proposed in [26]. In other words, instead of predicting the acoustic features related to the prosody of synthetic speech based on high-level linguistic features such as part-of-speech (POS) and semantic tags, more explicit control over the prosody of synthetic speech is assumed.

The corpus of American English includes recordings of 10 female and 13 male voices, with 9 of them including speech data in more than one speaking style. On the whole, there are 82 different SSCs, ranging in length from 1.1 minutes to 1 hour, with a median of 13.8 minutes. The recordings of 17 speakers have been made in a professional studio, while the recordings of the remaining 6 speakers have been obtained from public speeches available on the Internet, and their acoustic quality is inferior. The corpus of Mexican Spanish includes recordings of 54 female and 57 male voices, however, just two of them include more than one speaking style. On the whole, there are 123 different SSCs, ranging in length from 0.8 minutes to 28.8 minutes, with a median of 1.2 minutes. Only 2 voices, the ones including multiple speaking styles, were recorded in a professional studio, while the others have been obtained from different publicly available speech corpora, and their acoustic quality is inferior. The speech rates of the two languages are also significantly different (72 ms per phone for Spanish and 117 ms per phone for English). An overview of the speech data used in this research is given in Table I.

TABLE I. Speech Data Used in the Experiment

		American English	Mexican Spanish
Number of speakers	Female	10	54
	Male	13	57
	Total	**23**	**111**
Speaker/ style/cluster combinations (SSC)	Total number	82	123
	Min. duration	0:01:10	0:00:50
	Max. duration	0:59:59	0:28:49
	Mean duration	0:18:58	0:02:32
	Median duration	0:13:48	0:01:13
Duration	Studio quality	22:47:39	2:48:05
	Inferior quality	3:07:04	2:24:19
	Total	**25:54:43**	**5:12:24**

C. Model Implementation

The experiments are based on two single language models (for American English and Mexican Spanish) as well as one multilanguage model. The implementation of all models used in the experiments is based on the Merlin toolkit [27] and the TensorFlow framework [28].

1. Single Language Models

The single language models used for reference in this research represent monolingual TTS systems based on embedding (MTTSE), whose architecture is described in Section II.A.2.

The duration network has an input layer of size 641 for English and 610 for Spanish, 3 feedforward layers of size 1024 with tangent hyperbolic activation functions, one LSTM layer of size 1024, and a linear output layer of size 5, for the prediction of the durations of each HMM state of a phone. The exact size of the input layer for both languages is determined by their phoneme inventories and specific implementation details, which will be illustrated in detail for the case of American English. Firstly, besides the standard 39 phonemes in the phoneme inventory, the set of phoneme identifiers used in this research also included silence and non-phonemic glottal stop. Secondly, if the articulation of a phone was significantly impaired even taking into account its context, it was labelled as "damaged" and considered as a separate phoneme (e.g. "damaged /m/" as opposed to /m/). This was taken into account only in the input section related to the phonemic identity of the current phone, while in the sections related to positions ±1 and ±2 this impairment was disregarded. Thus, the size of the input section related to the current phone is 80, while the sizes of the remaining 4 sections are 41 each (Fig. 1). Finally, 82 SSC combinations and 315 prosodic features used for English bring the total size of the input layer to 641, as mentioned previously. Following the same reasoning, the total size of the input layer for Mexican Spanish (610) can be obtained taking into account the size of its standard phonemic inventory (28), the number of existing SSC combinations (123), and the number of prosodic features used (309). It should be noted that, although essentially the same prosodic annotation scheme was used for both languages, there were nevertheless certain language-specific features of minor significance, which explains the difference between the numbers of prosodic features used for the two languages. The sizes of SSC embeddings are 10 for the English model, and 12 for the Spanish model.

The architecture of the acoustic network is basically the same, with the input layer for both languages increased by 9 to accommodate for new frame related features [27]. As in the case of the duration network, hidden layers contain 1024 neurons, while the output layer contains 130 neurons, whose outputs correspond to the values of 40 mel-generalized cepstral coefficients (MGC), 2 band aperiodicity coefficients (BAP), the value of fundamental frequency, the first and second derivatives of all features previously mentioned, as well as one feature related to the degree of voicing (VUV).

2. Multilanguage Models

The multilanguage model used in the experiments has been built along the principles of cross-lingual TTS based on embedding (CTTSE), outlined in Section II.2.

In the multilanguage model a joint phoneme inventory of size 70 is used, including two non-phonemic glottal stops (one per language) as well as silence as phoneme identifiers. As was the case with MTTSE, the current phone is represented by a one-hot vector which considers poorly articulated phones as separate phonemes. Finally, 205 SSC combinations and 286 binary prosodic features shared between the two languages bring the total size of the input layer to 908 (Fig. 2). The size of the SSC embedding is set to 15, the size of the embedding related to the current phone is set to 10, while the sizes of the embeddings related to phones at positions ±1 and ±2 are set to 5.

3. The Choice of Hyperparameters

The choice of the size of the networks was largely based on our previous research related to embedding [22]. For instance, it has been shown that a smaller network (512 neurons per hidden layer instead of 1024) would produce synthetic speech of somewhat inferior quality. While most inputs to the networks are binary (0 or 1), the 9 frame-related inputs to the acoustic network are normalized to the range [0,1] at the global level. On the other hand, the output acoustic features are standardized (mean = 0, std = 1) at the level of an individual speaker, since it has been shown that doing otherwise would greatly degrade the quality of synthetic speech in a cross-lingual scenario. In cases of both MTTSE and CTTSE, the optimizer used is stochastic gradient descent with a momentum of 0.9, and starting learning rates of 0.008 for the duration network and 0.01 for the acoustic network.

As to the choice of the sizes of particular embeddings, from a theoretical standpoint, in order to keep the volume of a hypersphere which corresponds to one SSC (or one phoneme) in the embedding space relatively constant, a logarithmic dependency between the number of SSC (or the number of phonemes) and the embedding size is implied. However, in practice the choice of the size of an embedding is complicated by a number of issues. For instance, for an SSC embedding, it is not the same if a new SSC is actually a new speaker or just a new style of an existing speaker or even just a new cluster. In the research presented in [22], varying the embedding size from 4 to 40 for 67 SSC and a single language was shown to have surprisingly little impact on the performance. In this research, the size of the SSC embedding for the single language model for English was set to 10 (the values 10, 15 and 25 have been tested). Having in mind that there are about twice as many SSCs in the Spanish data, but that they also contain many more unique speakers, the size of the SSC embedding for the single language model for Spanish was set to 12. As to the size of phoneme embeddings in the multilanguage model, they greatly depend on the actual overlap between phonemic inventories of the two languages, not only on the phonological level, but on the phonetic level as well. Since the union of phoneme inventories of English and Spanish contains 67 phonemes, and since in the case of the current phoneme a phone with impaired articulation was considered as a separate phone, the values of 5, 10 and 15 were tested as the sizes of the embedding of the phonemic identity of the current phone, and the value of 10 was chosen as the one producing the highest quality of synthetic speech. In case of phones at positions ±1 and ±2, the size of the embedding was set to 5 since they carry less important information, and the impairment in their articulation is disregarded (i.e. impaired phones are not treated as separate phonemes). It was also established that, although final synthesis does not vary much in quality, embedding space looks more sensible for specific embedding sizes. Table II illustrates the case when the embedding size is set to 10, listing the nearest neighbours for certain phonemes. It can be seen, with some exceptions, that the distance in the embedding space indeed reflects the acoustic similarity between phonemes. For instance, English and Spanish /k/ are quite similar on the phonetic level as well, which is why the network has set them closely together in the acoustic space, unlike English and Spanish /b/, whose phonetic features are somewhat different. The anomaly of English /ʔ/ being identified as the closest neighbour of Spanish /a/ remains unexplained, but it should be noted that its influence on the quality of synthesis may be minor, since the acoustic features are formed not only on the basis of the current phoneme embedding but on the basis of embeddings at positions ±1 and ±2 as well, and /ʔ/, unlike /a/, is almost exclusively found between vowels. It should also be noted that the positions of embedding of phonemes with impaired articulation in the embedding space are irrelevant since these phonemes will never be used in synthesis.

TABLE II. Nearest Neighbours of Certain Phonemes in Case the Size of the Phoneme Identity Embedding Is Set to N = 10, With Respective Euclidean Distances Given in Brackets

Phone	1st neighbour	2nd neighbour	3rd neighbour
Sp. /b/	En. /w/ (2.61)	Sp. /w/ (3.48)	Sp. /g/ (3.49)
Sp. /k/	En. /k/ (3.04)	Sp. /g/ (3.21)	En. /g/ (3.43)
Sp. /r/	En. /r/ (4.20)	En. /ɚ/ (4.39)	En. /d/ (4.98)
Sp. /a/	En. /ʔ/ (2.49)	En. /ɑː/ (3.14)	En. /e/ (3.51)
Sp. /e/	En. /j/ (2.07)	En. /e/ (2.68)	En. /ɪ/ (2.16)
Sp. /u/	Sp. /o/ (2.41)	En./oʊ/ (3.03)	Sp. /w/ (3.41)

In both single language and multilanguage models, the duration network was trained for 100 epochs, while the acoustic network was trained for 45 epochs. Particular attention has been giving to the choice of the batch size. As the duration model is phone aligned, the batch size is represented as a product of the number of streams and the number of phonemes, where a single stream is made of concatenated sentences from the corpus. The batch size for the duration model was set to 8×50, which means that the update of weights is carried out each time a sequence of 50 phonemes from 8 different streams of sentences is processed by the network. The values given above were chosen after testing 4 to 16 streams and 16 to 50 phonemes per stream or even a single sentence as the entire batch, having in mind that for both languages the average sentence length in the corpus is close to 50 phonemes. Although it has been shown that the choice of one sentence per batch is satisfactory for synthesis of a speaker-language combination that exists in the training corpus, it is not suitable for cross-lingual scenario since it results in synthetic speech whose dynamics resemble the original language too much (e.g. an English speaker would speak Spanish too slowly). On the other hand, a batch size of 8×50 has shown to be suitable for high-quality synthesis regardless of whether the speaker-language combination exists in the training corpus or not. In the case of the acoustic network, batch size is represented as the product of the number of streams and the number of frames of length 5 ms per stream. By testing different combinations of values it has been found that, although high-quality synthesis for a speaker-language combination existing in the corpus is possible with batches as small as 32×25, cross-lingual scenario requires at least a batch size of 4×400. This implies that the update of weights should be done each time a sequence of 400 frames (corresponding to 2 seconds of speech, i.e. one half of an utterance of average length) from 4 different streams of sentences is processed by the network.

The imbalance between the representation of particular SSCs in the training corpora for both languages has been mitigated by using SSC-specific weight coefficients. Namely, when the total loss $J(\boldsymbol{\theta})_b$ for a batch is calculated, weight coefficients which boost the contributions of SSCs underrepresented in a particular training corpus are taken into account:

$$J(\boldsymbol{\theta})_b = \frac{1}{N_b}\sum_{j=1}^{N_b} w_j \sum_{i=1}^{N_{out}} \left(y_{ij} - t_{ij}\right)^2 \quad (1)$$

where N_b is the size of a batch (in samples, i.e. phones or frames), N_{out} is the size of the output layer, w_j is the weight coefficient corresponding to the SSC relevant to the j-th sample of the batch, and y_{ij} and t_{ij} are the calculated (predicted) and the target value of the i-th output for the j-th sample of the batch, respectively. The weight coefficient w_k corresponding to k-th SSC is given by:

$$w_k = \alpha\sqrt{N_k} \tag{2}$$

where N_k is the total number of utterances corresponding to k-th SSC, and α is a normalization factor given by:

$$\alpha = \sum_{k=1}^{N_{SSC}} \sqrt{N_k} \tag{3}$$

where N_{ssc} is the total number of SSCs.

4. Generation of Speech Waveforms

The first approach to the generation of speech waveforms from predicted acoustic features was based on WORLD, a widely-used deterministic vocoder [29]. It assumes a minimum phase for the spectrum and by using the predicted acoustic features it converts the cepstral features into a linear amplitude spectrum and produces excitation signal by mixing a pulse and a noise signal in the frequency domain, where each frequency band is weighted by a value of predicted band aperiodicity acoustic features. Finally, it generates a speech waveform based on the source-filter model. The second approach was based on WaveRNN [30], an increasingly popular neural vocoder, which predicts the more significant and the less significant halves of the 16-bit output sample separately, and supports simultaneous prediction of several output samples. Since it requires extreme processing power and large quantities of training data per speaker, it has been tested just for a single speaker who was most represented in the available training data, in order to establish whether the cross-lingual scenario is possible with a neural vocoder and how the use of a neural vocoder instead of a conventional one affects the quality of cross-lingual synthesis.

III. Experiments and Results

A. Experiment 1: Single Language Vs. Multilanguage Model

The aim of Experiment 1 is to compare the quality of speech generated using the multilanguage (ML) model and the single language (SL) model in a speaker's native language. Although ML in this experiment supports only two languages, it can be easily extended to more languages. It should be emphasized that in this research an extreme disbalance between the corpora of two languages exists – the English corpus is 5 times bigger than the Spanish one (~25h vs ~5h) but includes far fewer different speakers (23 vs 111, or 83 unique SSC vs 124 unique SSC).

Since there are original and synthesized recordings of the same sentences (withheld during training) it was possible to conduct objective evaluation of the quality of synthesized speech, based on a comparison of values of acoustic features extracted from original recordings and those predicted by the TTS model. The standard measures are: mel-cepstral distance (MCD), root mean square error of the fundamental frequency (RMSE F0), correlation between predicted and true fundamental frequency (CORR F0), root mean square error of phoneme duration expressed in frames per phone (RMSE DUR) as well as correlation between predicted and true phoneme durations (CORR DUR). Table III shows the objective measures for each language.

Since subjective evaluation is still considered in the literature as the most reliable way of establishing the quality of speech synthesis, a comparison between synthetic speech obtained by ML and SL models was also carried out through a preference test including 31 non-native listeners, who declared themselves as having sound knowledge of both English and Spanish. Each listener was given 20 tasks (10 per language). In each task there were two sentences with the same linguistic content – one sentence synthesized by SL model and the other one by ML model. In the preference test each language was represented by 5 speakers, 2 male and 3 female ones. The listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer "no preference" was also acceptable. The results of the preference test are given in Fig. 3.

TABLE III. Objective Measures of Distance Between Synthesized and Natural Speech

		MCD (dB)	RMSE F0 (Hz)	CORR F0	RMSE DUR	CORR DUR
English	SL	5.26	32.30	0.90	5.79	0.84
	ML	5.39	33.34	0.89	5.58	0.85
Spanish	SL	5.29	24.39	0.91	5.68	0.77
	ML	5.19	24.39	0.91	5.61	0.78

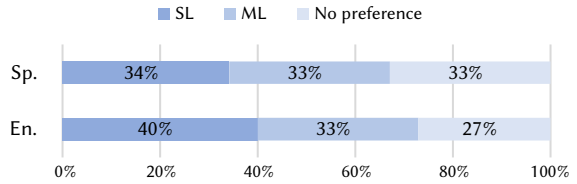


Fig. 3. Results of subjective comparison of the quality of speech synthesized by SL and ML.

B. Experiment 2: Speech Quality in a Cross-Lingual Scenario

The aim of Experiment 2 was to evaluate the quality of cross-lingual speech synthesis. Since in a cross-lingual scenario, ground truth examples (original recordings) do not exist, the only way of testing the quality of synthesis is subjective evaluation. Since it included rating the utterances delivered by different target speakers on a MOS scale rather than simple comparison, only native speakers of English and Spanish participated in the listening tests.

Two listening tests were carried out with 2 groups of 21 listeners per group – one for English and the other for Spanish. In each test there were 10 tasks, containing 4 utterances each. In each task the content of all utterances was the same, but two were synthesized by a speaker-language combination that exists in the training corpus, while the other two were synthesized by a speaker-language combination that does not exist in the corpus (i.e. cross-lingual scenario). Each of the 4 utterances in a task was delivered in the voice of a different speaker, of whom 2 were native English (male and female) and 2 native Spanish (male and female). In the entire test containing 10 tasks, there are sentences from 8 different speakers. In each task, listeners were asked to evaluate the quality of 4 synthesized sentences in terms of intelligibility and naturalness on a 1 to 5 MOS scale. Multiple speakers were introduced to neutralize any bias that a listener may have towards a specific voice. The results of the experiment are presented in Fig. 4. and Fig. 5.

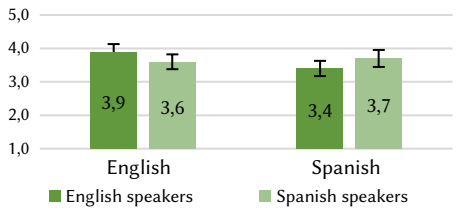


Fig. 4. Results of subjective comparison of the quality of original-language and cross-lingual synthesis (mean values with 95% confidence intervals are shown).

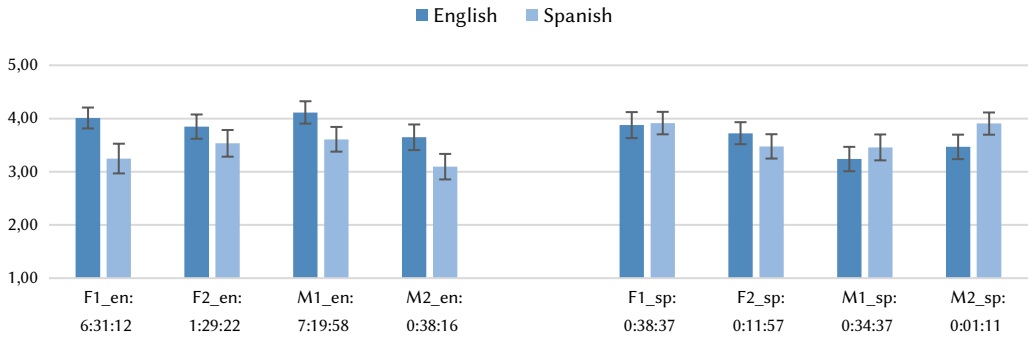


Fig. 5. Results of subjective comparison of the quality of original-language and cross-lingual synthesis for individual speakers (suffixes 'en' and 'sp' indicate the original language of each speaker). The amounts of available training data are also indicated.

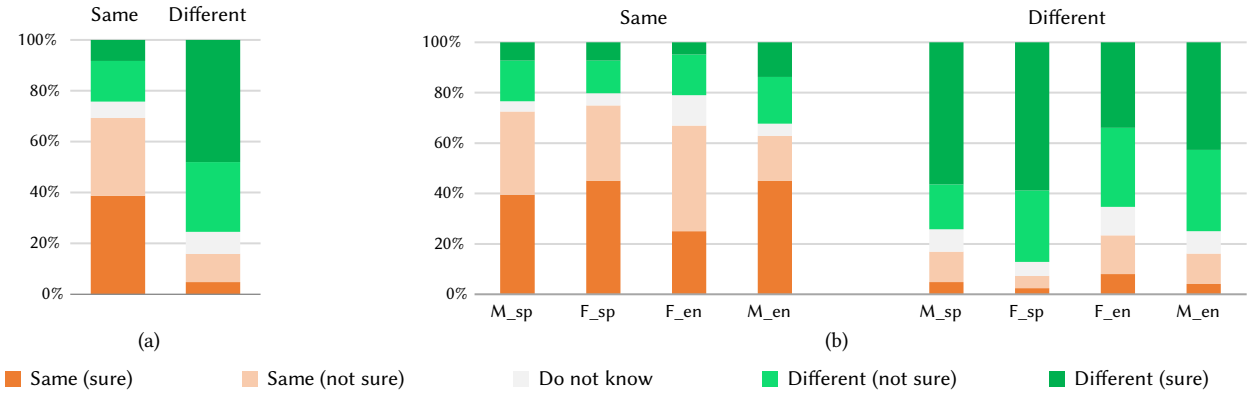


Fig. 6. Results of the evaluation of voice similarity in case of cross-lingual speech synthesis: (a) overall; (b) for each target speaker individually. Labels 'Same' and 'Different' indicate whether both sentences in a pair were delivered by the same speaker.

C. Experiment 3: Voice Similarity in a Cross-Lingual Scenario

The aim of Experiment 3 was to establish to what extent the characteristics of a speaker's voice remain preserved in the cross-lingual scenario. However, the task of evaluating voice characteristics is not easy when the entire sentence, even the language, is different. For that reason, the test has created as follows.

Each of the 32 tasks in a test included a pair of utterances and the listeners were asked to state their opinion as to whether both utterances were delivered by the same "virtual speaker" on a 1 to 5 scale defined as follows:

1. I am sure the utterances were delivered by different speakers;
2. I think they were delivered by different speakers;
3. I do not know whether they were delivered by the same speaker;
4. I think they were delivered by the same speaker;
5. I am sure they were delivered by the same speaker.

In each pair, one utterance was in Spanish and the other in English, both produced by the ML model. Consequently, in case of pairs of sentences from the same speaker, one sentence is necessarily synthesized using the cross-lingual scenario. There were 8 tasks for each of the 4 different target speakers (one for each combination of gender and native language). Half of the pairs of sentences in 32 tasks were delivered by the same speaker, while in the other half the utterances were delivered by different, but similar speakers. The test was presented to 31 non-native listeners, and the results are shown in Fig. 6.

D. Experiment 4: Adaptation to a New Speaker

The aim of Experiment 4 is to establish whether it is necessary to retrain the entire multispeaker (MS) ML model when a new speaker appears in order to obtain cross-lingual synthesis, or it is sufficient to adapt the existing MS ML model to new speaker data, as described in Section II.A.3. In the experiment two new native English speakers were introduced – a female, whose training corpus can be considered as small (10 min, 4 SSCs, inferior quality), and a male, whose training corpus can be considered as being of moderate size (45 min, 3 SSCs, studio quality).

The preference test consisted of 20 pairs of utterances, 10 per speaker, of which 5 were in English (original speaker-language combination) and 5 in Spanish (cross-lingual scenario). In each pair of utterances, one was synthesized by the multispeaker model that included the target speaker in the training corpus, while the other was synthesized by the model which had been adapted to the target speaker. A total of 31 non-native listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer "no preference" was also acceptable. The results are shown in Fig. 7.

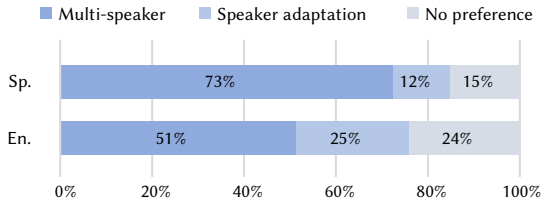


Fig. 7. Results of subjective comparison of the quality of multi speaker and speaker adaptation synthesis.

E. Experiment 5: Deterministic Vs. Neural Vocoder

The aim of Experiment 5 is to compare results that can be obtained by using a deterministic vocoder and from the data-driven vocoder, both in original-language and cross-lingual scenario. In this research

WORLD vocoder was used as a widely-used example of a deterministic vocoder, while WaveRNN was selected as an increasingly popular and efficient data-driven vocoder. At first sight, the experiment may not seem immediately related to this research since any vocoder produces speech samples given appropriate acoustic features at its input, and thus it may not be obvious that the purpose of the experiment is anything beyond a simple comparison between WORLD and WaveRNN. However, it should be taken into account that a neural vocoder has to be trained in order to produce speech based on acoustic features, and in the cross-lingual scenario, training data for a specific voice in the target language does not exist. In other words, the vocoder is required to produce speech in a language it has not heard before, and the experiment aims to establish whether this is at all practicable, and whether the possible loss in the quality of synthesized speech is acceptable.

A comparison between synthetic speech obtained by WORLD and WaveRNN was carried out through a preference test including 31 non-native listeners, who declared themselves as having sound knowledge of both English and Spanish. Each listener was given 10 tasks (5 per language). In each task the listeners were presented with two sentences with the same linguistic content – one sentence synthesized using WORLD and the other one using WaveRNN. All speech samples used in the experiment were synthesized in the voice of the English speaker (a female one, neutral style) with the most data in the training corpus. The samples representing original speaker-language combinations and cross-lingual synthesis were equally represented and randomly distributed in the test. The listeners were asked to select the utterance of better quality in terms of intelligibility and naturalness, and the answer "no preference" was also acceptable. The results of Experiment 5 are presented in Fig 8.

IV. Discussion

In this section an interpretation of the results will be presented, outlining their importance in view of the recognized limitations of the study. To illustrate the quality of speech synthesized by different approaches described in the paper, and to substantiate the results of the listening tests, the speech samples used in the tests have been made available at: www.alfanum.ftn.uns.ac.rs/crosslingual.

A. Experiment 1: Single Language Vs. Multilanguage Model

Based on objective measures of distance between acoustic features of original and synthesized speech (Table III), it can be concluded that in most cases the performance in a certain language of the ML model matches the performance of SL models. The results related to phone durations are slightly improved, as well as acoustic measures for the language with less data. This is in line with the expectation that, given enough training data, a SL model is capable of producing synthetic speech of high quality, and the use of a ML model may be advantageous since it is capable of overcoming the lack of training data for underrepresented languages. On the other hand, the results of the subjective evaluation (Fig. 3) show that there is no significant difference between compared methods, although the SL model is slightly preferable than the ML model in the case of English. This can be explained by the fact that in case of English much more training data were available, and since the SL model had enough material for training, adding a new language was only a distraction for the network. On the other hand, in case of Spanish, a three-way tie among the preference of SL, preference of ML and no preference at all, indicates that the additional language was not a distraction, although it did not improve the Spanish synthesis either. Our expectation that the benefit from the use of ML model increases with the increase of the number of underrepresented languages as well as the scarcity of training data, will be the subject of further research, as soon as prosodically annotated data in more languages become available.

It should be noted that a similar analysis has been conducted in [18], including a comparison of the quality of synthesis from end-to-end SL and ML models. Speech data used in [18] also exhibited a significant imbalance regarding the representation of each language (387 hours of English vs. 97 hours of Spanish and 68 hours of Chinese). The ML model was based on three languages – English, Spanish and Chinese. It was concluded that there was no significant degradation in case of ML, although SL was found to be slightly preferable (MOS grades were 0.1 lower for ML on average). It was also confirmed that SL was especially preferable in case of English (the difference in grades was more than 0.2), than in case of other two languages (the difference in grades was less than 0.1).

B. Experiment 2: Speech Quality in a Cross-Lingual Scenario

From the results of Experiment 2, shown in Fig. 4 and Fig. 5, it can be seen that, in general, higher quality synthesis is achieved for English, which could be expected since the English training corpus is 5 times larger than the Spanish one. However, taking into account only original speaker-language combinations (i.e. cases where speech is synthesized in the original language of the speaker), it is interesting to note that synthesis of Spanish speech was, on average, rated only 0.2 lower than synthesis of English speech. This is a very encouraging result, having in mind the difference in the sizes of training corpora. More importantly, the results of Experiment 2 fully confirm the feasibility of cross-lingual speech synthesis based on CTTSE. Namely, the quality of cross-lingual synthesis has been rated as inferior to original-language synthesis by only 0.3 in both cases (English to Spanish and vice versa).

Analysing grades obtained for each individual speaker, it can be noticed that all target English speakers have grades higher than 3.5 for synthesis in English, 2 of them even grades higher than 4.0, while the target Spanish speakers have obtained lower grades for synthesis in Spanish, 2 of them even less than 3.5. This may be explained by the fact that all English speakers have much larger training corpora than their Spanish counterparts. However, it is interesting to note that the Spanish speaker with only 1 minute of training data and the English speaker with more than 7 hours of training data obtained grades which differ by only 0.21 for their native languages. The differences between synthesis in the original language and cross-lingual scenario are smaller (up to 0.45) in case of Spanish speakers, where in one case the cross-lingual scenario was graded even better than synthesis in the original language. For English speakers, the cross-lingual scenario is graded usually as at least half a grade inferior to synthesis in the original language. However, no speaker-language combination obtained a grade below 3.1.

Although it was not mentioned in the comments of listeners since they were unaware of the origin of each utterance, it can safely be concluded that inadequate speech rate is one of the factors that have reduced the quality of cross-lingual speech synthesis. By comparing cross-lingual speech samples, it can be concluded that synthesized voices tend to preserve the speech rate of their original language, which implies that English voices speak Spanish unnaturally slow, while Spanish voices speak English unnaturally fast. This may be the consequence of the specific approach to network output normalization, which will be investigated further.

The quality of synthesis for underrepresented languages in [18] was graded as equal or slightly inferior to the case of the language which was represented with the most training data for an original speaker-language combination (e.g. for Chinese the average MOS grade was approximately 0.3 lower than for English and Spanish). Switching to a cross-lingual scenario introduced a slight degradation in quality

(0.06 lower MOS grade in case of 3 languages, and 0.13 in case of only Spanish to English and vice versa). Although these results are better than in our research, it should be noted that in our research the amount of training data was 15 to 20 times smaller. Another research, presented in [13], contrary to our research and [18], used corpora of bilingual speakers in order to construct a ML model. They also conducted an experiment with 2 bilingual speakers and 1 monolingual speaker and tested the cross-lingual scenario (about 45 minutes of data for each speaker-language combination was used). The quality of synthesis obtained by the ML model in a cross-lingual scenario was graded with a MOS grade by 0.25 lower than in case of a SL model trained on data from only one speaker (standard TTS).

C. Experiment 3: Voice Similarity in a Cross-Lingual Scenario

Experiment 3 aimed at establishing to what degree speech synthesized in another language retains the voice characteristics of the original speaker. As explained in Section III.C, the participants were asked to state whether they believe that each of the two utterances presented in a pair was delivered by the same "virtual speaker" on a 1 to 5 scale. If grades 5 and 4 can be considered as correct answers and grades 1 and 2 as wrong answers for pairs where the utterances correspond to the same speaker, and vice versa if they correspond to different speakers, the listeners answered correctly in 72% cases, could not decide in 8%, and gave the wrong answer in 20%. Fig. 6. provides a more detailed analysis of the results. Since the listeners recognized correctly that the speaker was the same in almost 70% cases, being sure in their answers (grade 5) in almost 40% cases, it can be concluded that the voice characteristics remained preserved in cross-lingual scenario. On the other hand, in case when the sentences in the pairs were actually delivered by different speakers, listeners correctly recognized it in almost 80% of cases, being sure (grade 1) in almost 50%. It can be noted that for the female English speaker the listeners were less sure in their answers and also the most indecisive in pairs where her voice was present.

It should be noted that the reported degradation in voice similarity in [18] in comparison with the original speaker-language case was as high as 1.0. The authors of [18] have also emphasized the problem of grading voice similarity in case the sentence or even the language is different, which is why we have opted for a different approach – to ask the listeners to identify whether the two utterances in different languages have been delivered by the same speaker. On the other hand, the evaluation of the voice similarity in [13] was quite simple. Namely, owing to the use of bilingual speakers, it was possible to directly compare the result of synthesis from the ML model in a cross-lingual scenario with an original recording of the speaker in the target language. A decrease of the MOS grade by 0.58 with respect to the synthesis from SL model was reported.

D. Experiment 4: Adaptation to a New Speaker

The results of Experiment 4, shown in Fig. 7, indicate that re-training the entire MS ML model from scratch including the new speaker produces speech of better quality than speaker adaptation (SA). The preference of MS ML over the SA approach is more emphasized in the case of Spanish, i.e. in case of the cross-lingual scenario. It can be assumed that, in adapting the existing model to the new speaker, the network is less ready to generalize and produce a new speaker-language combination because it overfits to the single speaker-language combination used for adaptation.

The results do not differ much depending on whether the training corpus is small or of moderate size, although SA has shown to be more acceptable in the case of the speaker with a moderate training corpus. It is interesting to note that during speaker adaptation, embedding values for the phonemes of the language not existing in the corpus of the new speaker will not be updated. However, this should not lead to a significant difference in quality with respect to the cross-lingual scenario in which a speaker is included in the training of the original ML model, since in that case his/her data will influence only the embeddings for the phonemes of languages that exist in his/her training corpus.

E. Experiment 5: Deterministic Vs. Neural Vocoder

From the results of Experiment 5, shown in Fig. 8, it can be seen that, while WaveRNN is preferable over WORLD in both English and Spanish, i.e. in both original language and cross-lingual synthesis, the preference in case of English is negligible. As is well known, both vocoders have their own specific properties, e.g. while synthesis by WORLD is relatively stable but with a constant impairment in quality referred to as "buzzing" [31], the synthesis by WaveRNN generally sounds more natural but is less stable. A point of some relevance for this research is that WaveRNN synthesis includes a certain overtone which may affect the timbre of the voice, but it could not be spotted by listeners to whom the original voice is unknown. It should also be noted that, unlike WORLD, WaveRNN exhibits significant flexibility in terms of architecture and hyperparameters, so the results can be further improved. However, a downside of WaveRNN is the necessity of large corpora, and in this experiment, only one speaker with the sizable corpus was used, so its adaptation or multispeaker training are the subjects of further research.

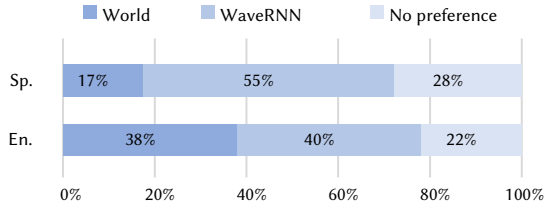


Fig. 8. Results of subjective comparison of the quality of utterances obtained by using WORLD and WaveRNN vocoders.

Most importantly, the experiment has shown that WaveRNN is able to produce high-quality synthetic speech even in the language it is not initially trained on, which confirms the assumption that acoustic features from the TTS model of one speaker are meaningful input to WaveRNN regardless of the language to which they may correspond.

V. Conclusion

The study presents a novel method for multilingual and cross-lingual neural network speech synthesis. Firstly, it shows that the proposed method is capable of speech synthesis in multiple languages, and that it is a good basis for the creation of speech synthesis for languages in which a relatively small quantity of speech data is available. As its main point, the study shows that it is possible to synthesize speech in a specific person's voice in a language that this person has never spoken. The quality of cross-lingual synthesis in terms of intelligibility and naturalness, as well as the resemblance of the synthesized voice in a cross-lingual scenario to the same voice in original language synthesis, were both established to be relatively high (a difference in quality on a MOS scale was found to be 0.3). Since it would be impractical to retrain the entire system each time a new speaker is introduced, a method for speaker adaptation in a cross-lingual scenario was examined as an alternative and it was found that does not lead to an unacceptable loss in speech quality, particularly in the case of the language with greater overall quantity of training data. Finally, it has been shown that the proposed method for cross-lingual synthesis supports the use of neural vocoders, even though it

means that they have to be trained on data in one language, and used for synthesis of speech in another. The study, thus, brings the state of the art in speech technology one step closer to the synthesis of arbitrary text in an arbitrary voice, speaking style and language, easily extensible to new speakers, styles and languages.

The study is somewhat limited by the fact that it was based on only two languages, with significant differences in both the number of speakers in the training corpus as well as the average quantity of available data per speaker. However, most of its results and conclusions are in agreement with expectations based on theoretical knowledge. Our future research on this topic will include the extension of the study to multiple languages as soon as more data become available. The study also raises a number of questions related to specific implementation of particular models, most notably the normalization of network outputs, which will also be investigated further in our future work.

Acknowledgment

This research was supported by Speech Morphing Systems Inc., Campbell, CA, United States of America, as well as the Science Fund of the Republic of Serbia (grant #6524560, AI – S-ADAPT). Speech corpora used in the research were provided by Speech Morphing Systems Inc. for research purposes.

References

[1] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH 1999*, Budapest, Hungary, 1999, pp. 835–838.

[2] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft Mulan – a bilingual TTS system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2003*, Hong Kong, China, 2003, vol. I, pp. 264–267.

[3] N. Campbell, "Foreign-language speech synthesis," in *Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis,* Jenolan Caves, Australia, 1998, pp. 177–180.

[4] M. Moberg, K. Pärssinen, and J. Iso Sipilä, "Cross-lingual phoneme mapping for multilingual synthesis systems," in *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP 2004,* Jeju Island, Korea, 2004, pp. 1029–1032.

[5] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign TTS," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, USA, 2004, pp. 217–218.

[6] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2011*, Prague, Czech Republic, pp. 5120-5123.

[7] J. He, Y. Qian, and F. K. Soong, "Turning a monolingual speaker into multilingual for a mixed-language TTS," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association INTERSPEECH 2012*, Portland, OR, USA, 2012, pp. 963–966.

[8] Y. Qian, H. Liang and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin–English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009, doi: 10.1109/TASL.2009.2015708.

[9] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*, Brighton, United Kingdom, 2009, pp. 528–531.

[10] H. Zen et al., "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012, doi: 10.1109/TASL.2012.2187195.

[11] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association INTERSPEECH 2016,* San Francisco, CA, USA, 2016, pp. 322–326.

[12] F. L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5515–5519.

[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5540–5544.

[14] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2016*, Shanghai, China, 2016, pp. 5135–5139.

[15] J. Sotelo et al., "Char2wav: End-to-end speech synthesis," in *Proceedings of the 5th International Conference on Learning Representations ICLR 2017,* Toulon, France, 2017, pp. 1–6.

[16] Y. Wang et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," arXiv preprint arXiv:1703.10135, 2017. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1703.10135.

[17] S.Ö. Arık et al., "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning PMLR*, Sydney, Australia, 2017, vol. 70, pp. 195–204.

[18] Y. Zhang et al., "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," arXiv preprint arXiv:1907.04448, 2019. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1907.04448.

[19] M. Chen et al., "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding", in *Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019,* Graz, Austria, 2019, pp. 2105–2109.

[20] Z. Liu and B. Mak, "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers," arXiv preprint arXiv:1911.11601, 2019. Accessed: July 15, 2020. [Online]. Available: https://arxiv.org/abs/1911.11601.

[21] S. Maiti, E. Marchi, and A. Conkie, "Generating multilingual voices using speaker space translation based on bilingual speaker data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2020*, Shanghai, China, 2020, 7624–7628.

[22] M. Sečujski, D. Pekar, S. Suzić, A. Smirnov, and T. Nosek, "Speaker/style-dependent neural network speech synthesis based on speaker/style embedding", *Journal of Universal Computer Science*, vol. 26, no. 4, pp. 434–453, 2020.

[23] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association INTERSPEECH 2016,* San Francisco, CA, USA, 2016, pp. 2278–2282.

[24] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2015*, South Brisbane, Australia, pp. 4475-4479.

[25] T. Delić, S. Suzić, M. Sečujski, and D. Pekar, "Rapid development of new TTS voices by neural network adaptation," in *Proceedings of the 17th International Symposium INFOTEH-JAHORINA,* Jahorina, Bosnia and Herzegovina, 2018, pp. 1–6.

[26] M. Sečujski, S. Suzić, S. Ostrogonac, and D. Pekar, "Learning prosodic stress from data in neural network based text-to-speech synthesis," *SPIIRAS Proceedings*, vol. 4, no. 59, pp. 192–215, 2018, doi: 10.15622/sp.59.8

[27] Z. Wu, O. Watts, and S. King, "Merlin: an open source neural network speech synthesis system", in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, 2016, pp. 218–223.

[28] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation OSDI 2016*, Savannah, GA, USA, 2016, pp. 265-283.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp.1877–1884, 2016, doi: 10.1587/transinf.2015EDP7457.

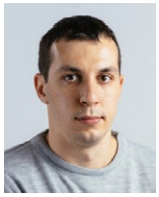
[30] N. Kalchbrenner et al., "Efficient neural audio synthesis," arXiv preprint

arXiv:1802.08435, 2018. Accessed: July 18, 2020. [Online]. Available: https://arxiv.org/abs/1802.08435.

[31] S. King, "An Introduction to Statistical Parametric Speech Synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

Tijana Nosek

Tijana Nosek, MSc, born in 1992, is a doctoral candidate at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, Chair of Telecommunications and Signal Processing. She is engaged as teaching assistant at the Faculty of Technical Sciences, at courses principally related to machine learning and digital signal processing. She has participated in a number of national and international scientific projects and currently is a member of the team which participates in the scientific project Speaker/Style Adaptation for Digital Voice Assistants Based on Image Processing Methods (S-ADAPT), financed by the Science Fund of the Republic of Serbia, with the aim of improving machine learning algorithms for speech synthesis with limited training data. She also cooperates closely with "AlfaNum – Speech Technologies", a leading company in speech technologies in the region. The main area of her research is speech synthesis based on neural networks with particular focus on expressive speech. She authored or co-authored 4 research papers in renowned international journals as well as 19 papers at scientific conferences related to speech processing. She is a member of international organizations such as IEEE and Audio Engineering Society (AES).



Siniša Suzić

Siniša Suzić, PhD, was born in 1988. He defended his bachelor, master and PhD thesis at the University of Novi Sad, Faculty of Technical Sciences, in 2011, 2012 and 2019, respectively. He is currently engaged as teaching assistant at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, at several courses related to acoustics and signal processing. He has participated in a number of national and international scientific projects related to speech processing and human-machine interaction. His areas of scientific interest include expressive speech synthesis as well as speech synthesis with limited training data. He also contributed to the creation of speech resources used in the development of large vocabulary speech recognition for Serbian and other South Slavic languages, as well as speech recognition for mobile phones. He also contributed to the development of different commercial, speech related software in Serbian and other South Slavic languages. He authored or co-authored 4 research papers in renowned international journals as well as more than 25 papers at scientific conferences related to speech processing. He is a member of IEEE.

Darko Pekar

Darko Pekar, MSc, was born in 1972. He graduated in 1998, at the University of Novi Sad, Faculty of Technical Sciences. Up to 2003 he was the leading expert at a university research project where he obtained wide-ranging experience in the field of speech technology, as well as management of scientific and technological projects. In 2003 he became CEO of the company "AlfaNum", and in cooperation with University of Novi Sad, he has continued to manage teams working on speech technology and machine learning. He currently leads the team of 25 software engineers and accompanying staff. His major achievements over the years include the development of: high quality text-to-speech systems for Serbian, Croatian, English, Spanish and Hebrew languages; large vocabulary speech recognition systems for Serbian and Croatian languages; methods for speaker adaptation by using very small quantities of speech data. Until 2019 he was also engaged as research assistant at the Faculty of Technical Sciences, and is currently finishing his PhD thesis in the area of speaker adaptation. Although he focuses on practical development and providing market-ready solutions, he has also published more than 100 articles and papers in national and international scientific journals.

Radovan Obradović

Radovan Obradović, MSc, was born in 1969 and in 1999 he received his BSc/MSc degree in Electrical Engineering from the University of Novi Sad, Faculty of Technical Sciences. He has worked in industry as a researcher in the fields of digital signal processing, speech recognition and synthesis, natural language processing and computer vision. During his cooperation with the company "AlfaNum", he has played a major role in the development of a range of speech technology solutions for Serbian and other languages, related to both speech recognition and synthesis. His current research interest includes neural speech synthesis, speech recognition, dialogue systems, applications of sparse representations in artificial neural networks, biologically inspired learning and meta learning.

Milan Sečujski

Milan Sečujski, PhD, was born in 1975, and currently works as Associate Professor at the University of Novi Sad, Faculty of Technical Sciences, Department for Power, Electronics and Telecommunication Engineering, Chair of Telecommunications and Signal Processing. He is engaged as a lecturer in university courses related to digital signal processing, time series analysis as well as machine learning. His areas of scientific interest include computational linguistics, speech and language technology, as well as human-machine interaction. The result of his master research thesis evolved into the highest quality speech synthesizer in Serbian. This system has since been widely used, initially by the blind and visually impaired computer users, but its use has since spread into the domain of telecommunications services, where it has remained the most widely used speech system in Serbian and Croatian. He has participated in a number of international projects, and is currently participating in the Erasmus+ project BENEFIT (Boosting the Telecommunications Engineer Profile to Meet Modern Society and Industry Needs). His current research includes natural language processing as well as mathematical modeling of the prosodic features of speech, most notably for the purposes of expressive speech synthesis as well as speaker conversion in speech synthesis. Apart from his work in the domain of speech technology, he has made scientific contribution in the field of acoustic metamaterials as well. He is a member of international organizations such as IEEE and Audio Engineering Society (AES).

Vlado Delić

Vlado Delić, PhD, was born in 1964. He is engaged as Full Professor at the University of Novi Sad, Faculty of Technical Sciences, Serbia, and he is also the head of the Chair of Communication Engineering and Signal Processing. He received the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Belgrade and University of Novi Sad, in 1993 and 1997, respectively. He has created distinguished curricula in acoustics, audio signal processing, and speech technologies at FTS-UNS, which have attracted interest of several universities from the region who invited prof. Delić as a visiting professor. He has been leading major national and international research proje cts in the field of speech technologies in Serbia, and is the leader of the most renowned scientific research team in the field of speech technology in Western Balkans. Prof. Delić has large experience and skills in signal processing research and transfer to ICT applications, he has published nearly 300 scientific papers, and his research has evolved into a number of technical solutions widely applied across the region. For his contribution to innovation in the field of speech technology, Prof. Delić has received several prestigious awards. He is a member of international organizations such as IEEE and Audio Engineering Society (AES).