ENHANCING EXPRESSIVITY TRANSFER IN TEXTLESS SPEECH-TO-SPEECH TRANSLATION

Jarod Duret, Benjamin O'Brien, Yannick Estève, Titouan Parcollet

ENHANCING EXPRESSIVITY TRANSFER IN TEXTLESS SPEECH-TO-SPEECH TRANSLATION

Jarod Duret[1], *Benjamin O'Brien*[1], *Yannick Estève*[1], *Titouan Parcollet*[2]

[1]LIA - Avignon Universite, France
[2]University of Cambridge, United-Kingdom

ABSTRACT

Textless speech-to-speech translation systems are rapidly advancing, thanks to the integration of self-supervised learning techniques. However, existing state-of-the-art systems fall short when it comes to capturing and transferring expressivity accurately across different languages. Expressivity plays a vital role in conveying emotions, nuances, and cultural subtleties, thereby enhancing communication across diverse languages. To address this issue this study presents a novel method that operates at the discrete speech unit level and leverages multilingual emotion embeddings to capture language-agnostic information. Specifically, we demonstrate how these embeddings can be used to effectively predict the pitch and duration of speech units in the target language. Through objective and subjective experiments conducted on a French-to-English translation task, our findings highlight the superior expressivity transfer achieved by our approach compared to current state-of-the-art systems.

Index Terms— speech translation, prosody prediction, speech generation

1. INTRODUCTION

In today's interconnected world, speech-to-speech translation (S2ST) technology can help bridge the communication gap between people speaking different languages by enabling effective communication across diverse languages and cultures. Nevertheless, existing speech-to-speech translation systems frequently fall short of retaining the subtleties of expressiveness embedded within the speaker's original message. Developing speech-to-speech translation systems capable of capturing the emotional and expressive dimensions of spoken language is crucial to improve the naturalness of speech generation. Conventional speech-to-speech translation systems rely on cascaded approaches [1, 2] that follow a two-step approach first converting the source speech into a textual representation in the target language domain. This can be accomplished by using automatic speech recognition (ASR) followed by machine translation (MT), or by using an end-to-end speech-to-text translation (S2T) system [3, 4]. The resulting text output is then transformed into a speech using text-to-speech (TTS).

More recently, a textless direct speech-to-speech translation (S2ST) approach has been proposed which relies on discrete speech units [5]. This approach is particularly valuable when translating from an unwritten language and/or to an unwritten language. Furthermore, it has been observed that this method is also highly effective for languages that possess a written form [5, 6]. This technique is designed to effectively capture the linguistic content of the target speech while minimizing the impact of the speaker's prosodic features. Previous study [7] has shown that the utilization of discrete speech units successfully disentangles linguistic content from the influence of prosodic characteristics and speaker identity.

Another challenge in preserving expressivity in speech-to-speech translation is the lack of parallel annotated speech data. The recent approach introduced in [6] is designed to address this lack of paired speech data by focusing on the linguistic context. However, this approach does not address the issues of preserving emotions and other non-linguistic information contained in the source language speech.

Inspired by a recent work on prosody reconstruction from multilingual speech representation [8], we propose an approach that aims to build a speech-to-speech translation system that preserves the expressivity without the need for parallel speech data. This approach consists of training a multilingual emotion embedding extractor used to compute an emotion embedding from an utterance and exploit it for speech resynthesis. In our work, we extend this approach to address emotion preservation in textless speech-to-speech translation. We compute an emotion embedding from the source utterance and use it to condition the duration and pitch predictor models used for generating the target utterance from a discrete speech unit representation.

2. RELATED WORK

Spoken Language Modeling from audio. Generative spoken language modeling from audio is a task that involves acquiring the acoustic and linguistic characteristics of a language solely from raw audio data, without any accompanying text or labeled information. In [9], the authors proposed to leverage advancements in self-supervised speech representation learning to discover discrete speech units and subse-

979-8-3503-0689-7/23/$31.00

quently use them in downstream tasks. They demonstrate that speech generation can be achieved by sampling sequences of these discovered units from a unit-discovery model and synthesizing them into a coherent speech waveform using a unit-to-speech model. Building upon this work, in [7], the authors demonstrated the effectiveness of utilizing self-supervised learned discrete speech units for generating high-quality speech. Furthermore, a comparable approach and speech representation scheme were employed for the purpose of textless speech emotion conversion through translation [10] and for prosody reconstruction using a multilingual speech representation [8]. In this study, we leverage a similar approach and speech representation scheme to encode target speech in order to train a speech-to-unit translation model.

Speech-to-speech translation. Most speech-to-speech translation systems rely on cascaded approaches that require intermediate text representation. This makes them unusable for languages without written forms or datasets containing only speech alignments. Recent research on S2ST is new, exploring scenarios involving speech-to-speech translation (S2ST) that does not rely on intermediate text representation. In [4], an attention-based sequence-to-sequence neural network was proposed to enable direct speech translation without the need for intermediate text representation. The model was trained end-to-end, mapping speech spectrograms from a source language to target spectrograms in another language. Additionally, the authors introduced a variation that aimed to transfer the voice characteristics of the source speaker to the translated speech. However, as the model was trained on synthetic data, the voice transfer capabilities did not achieve comparable results to those observed in a similar text-to-speech context. In subsequent work, [5] introduces a direct S2ST system based on self-supervised discrete representations. The proposed approach exhibits enhanced performance compared to its predecessor, unfortunately, it remains constrained by the utilization of synthetic data. Furthermore, it is important to note that this study did not emphasize the exploration of paralinguistic information. More recently, [6] tackles direct S2ST by following [5] and focuses on training the system with real-world data on multiple language pairs. Previous studies in the field of direct speech-to-speech translation have predominantly concentrated on improving the quality of the translation, disregarding the paralinguistic dimension and expressivity transfer. In contrast, the current study aims to build a speech-to-speech translation framework that can transfer the expressivity from one language into another.

3. ARCHITECTURE

Our speech-to-speech translation framework does not require parallel speech data for speaker and expressivity modeling, enables the translation of speech while maintaining the inherent expressive content, and can generate speech in the target language with multiple voices. The proposed framework can be decomposed into two parts. First, a speech-to-unit translation model (Section 3.1), composed of a speech encoder and an acoustic decoder. Secondly, a unit-to-speech synthesizer (Section 3.2), composed of an emotion encoder, a speaker encoder, a duration predictor, a pitch predictor and a speech vocoder. The following subsections describe each component of the proposed S2ST framework while the overall architecture is illustrated in Figure 1.

3.1. Speech-to-unit translation model

The following describes the speech-to-unit translation (S2UT) model (1) depicted in Figure 1. In order to capture the linguistic content, particularly pseudo-phonetic information present in speech, we employ a pre-trained self-supervised learning (SSL) model to extract raw speech features from the audio signal, namely multilingual HuBERT (mHuBERT) [6] for English and Wav2Vec 2.0 [11] for French. Wav2Vec 2.0 and mHuBERT models are pre-trained in a self-supervised manner and produce continuous representations for every 20-ms frame. To extract the sequence of speech units, a k-means clustering is applied to the raw speech features and the learned K cluster centroids are used to transform audio into a sequence of cluster indices at every 20ms of the input audio signal. For English speech, we extracted representations from the 11^{th} layer of mHuBERT model and set $k = 1000$ as used in [6] for speech-to-speech translation. For French speech, we extracted representations from the 11^{th} layer of Wav2Vec2-XLSR model and set $K = 1000$. Following [6][10], as a way of speeding up training and inference time, we experimented with reducing a sequence of units to a sequence of unique units by removing consecutive duplicated units (e.g., 0, 0, 1, 1, 1, 2 $\rightarrow$ 0, 1, 2). We denote such sequences as "reduced".

We build the S2UT model by adapting the transformer encoder-decoder framework presented in [12]. As an encoder, we chose a large Wav2Vec 2.0 pre-trained on 7.6K hours of French speech (1.8K Males / 1.0K Females / 4.8K unknown) [1]. In contrast to the encoder, the decoder consists of 6 transformer layers with a random initialization for each transformer decoder weight. To alleviate the mismatch between the length of the source speech and the reduced target units, we introduced an adaptor layer of a single 1-D convolutional layer with stride 2 between the encoder and the decoder. We combined the Wav2Vec 2.0 encoder, the adaptor along with the transformer decoder and we finetune the whole model end-to-end. Following[6], we explored an auto-encoding style auxiliary task by adding a separate transformer decoder as auxiliary task to help the model converge during training. This separate transformer consists of 3 transformer layers, which are trained to predict the discrete units sequence of the source speech as the target.

[1] https://huggingface.co/LeBenchmark/wav2vec2-FR-7K-large

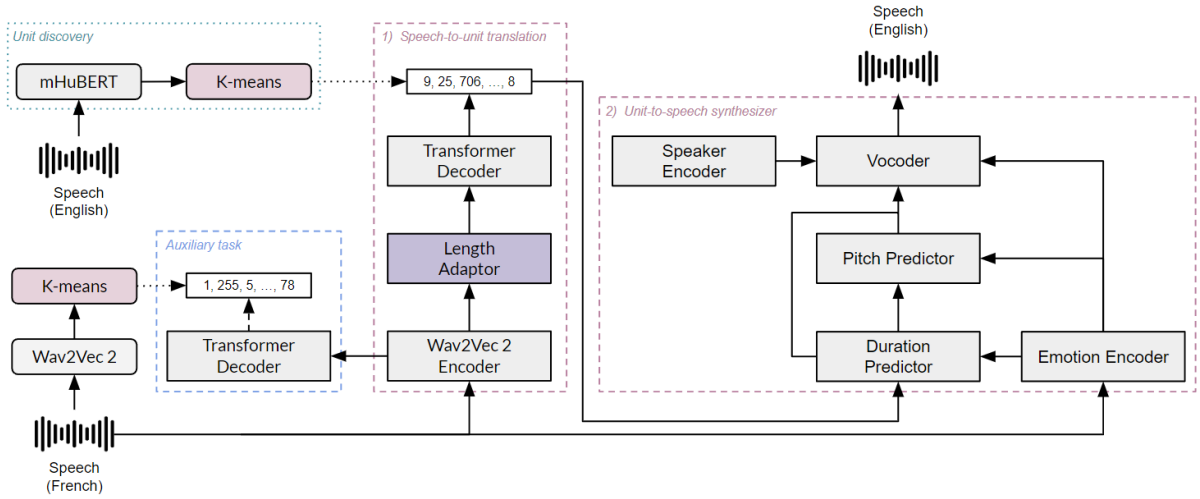

Fig. 1. Illustration of our proposed speech-to-speech translation model. First, the input speech is translated into a sequence of discrete units by the speech-to-unit translation model (1). Next, we predict duration and F0 before feeding them to a unit-to-speech model (2). Duration Predictor, Pitch Predictor, and the unit-to-speech model are conditioned by the emotion embedding extracted from the source speech by the emotion encoder. The speaker is encoded using a 1-hot vector directly in the unit-to-speech model.

3.2. Unit-to-speech model

The following section describes all components of the unit-to-speech (U2S) model (2) depicted in Figure 1.

Emotion Encoder To capture emotion representations, we use a dedicated encoder specifically trained for emotion recognition tasks within a multilingual context. Our proposed architecture is inspired by [8] and involves a pre-trained Wav2Vec2-XLSR encoder, a bottleneck layer, and a dense layer. We fine-tuned both the CNN and Transformer modules of the Wav2Vec2-XLSR model following the approach described in [13]. The information across the entire source speech sequence is encoded into a single fixed-length vector representation of size 96 by passing the audio through the bottleneck layer and applying temporal pooling.

Speaker Encoder In order to synthesize speech using the voices of several speakers, we introduce a speaker representation that serves as an additional conditioning factor. Inspired by [8], we optimize the parameters of a fixed-size look-up table. Although using speaker representations from a pre-trained speaker encoder enables generalization to new and unseen speakers, it is worth noting that such embedding captures a broader range of speaker-related information and it may be slightly less efficient in capturing solely the speaker identity, resulting in a degradation of synthesized speech.

Duration Predictor As the speech-to-unit translation model reduced sequences, we were required to predict the duration of each discrete unit before feeding them to the pitch predictor and unit-to-speech model. For this purpose, we take inspiration from work on TTS [14], where a CNN is used to predict the duration of each phoneme from a phoneme sequence. Following [10, 8], we replaced the phoneme sequence with the reduced discrete unit sequence and predict the number of repetitions for each unit, in order to reconstruct the original sequence. During training, we conditioned the model using an emotion embedding extracted from ground-truth speech and the ground-truth discrete unit duration is used as supervision. At inference time, the emotion embedding is extracted from the speech in the source language.

Pitch Predictor Pitch is an important characteristic of speech prosody, however, due to the non-monotonic alignment characteristic of speech-to-speech translation, a direct extraction of pitch from the source signal is not viable. Thus, an alternative method was required to accurately estimate pitch in this context. To overcome this limitation, we introduce a F0 estimation model to predict the pitch directly from a sequence of speech units. During the training phase, we use the ground-truth speech in order to extract the speech units sequence, and, during inference, we use the output of the S2UT model. Our pitch predictor model is a CNN followed by a linear layer projecting the output to $\mathbb{R}^d$. We apply a sigmoid on the model prediction to output a vector in $[0, 1]^d$. During the training phase, the target F0 is extracted using the YAAPT [15] algorithm. Following [10, 8], we discretize ranges of F0 values into d bins, represented by one-hot encodings. Then, we compute the weighted-average of the activated bins in order to expand the output range during the conversion of bins back to F0 values. We apply a normalization on the F0 values using the mean and standard deviation for each speaker.

Like the F0 estimation model, the duration predictor model is conditioned using an emotion embedding extracted from ground-truth speech. The same embedding is used to condition both models.

Speech synthesis Following [7], we use the HiFi-GAN neural vocoder [16] to synthesize speech. HiFiGAN is a generative adversarial network (GAN) that consists of one generator and a set of discriminators. The generator is a fully convolutional neural network. Inspired by [10], we adapted the generator architecture to take as input a sequence of discrete-unit inflated using the predicted durations, predicted F0, emotion-embedding, and a speaker-embedding. Before feeding the above features into the model, we concatenate them along the temporal axis. The sample rates of unit sequence and F0 are matched by means of linear interpolation, while the speaker-embedding and emotion-label are replicated along the temporal axis.

Regarding the set of discriminators, the model is composed of two modules: a Multi-Scale Discriminators (MSD) and a Multi-Period Discriminators (MPD). The first type operates on different sizes of sliding windows over the input signal, while the latter samples the signal at different periods.

4. EXPERIMENTAL SETUP

We use the SpeechMatrix [17] corpora for training and evaluating our speech-to-unit translation (S2UT) model. SpeechMatrix consists of 126 language pairs with a total of 418 thousand hours of speech from European Parliament recordings. In this study, only French-to-English language pairs were considered, yielding a $1,507$ hours train set.

In addition to the mined speech-to-speech data for training purposes, we extend our evaluation by leveraging labeled public speech datasets obtained from two distinct corpora that cover various domains. First, Europarl-ST (EPST) [18], a multilingual corpus containing paired audio-text samples built from recordings of debates from the European Parliament, containing 72 translation directions in 9 languages, including French to English direction. The second dataset is FLEURS [19]. Derived from from FLoRes [20], FLEURS is an extension that introduces speech recordings for these translated texts, resulting in a collection of speech-to-speech data comprising French to English direction. FLEURS texts are from English Wikipedia. During training, we extract a validation set from SpeechMatrix of about 1000 sample which are not in the test set. FLEURS validation set is derived from its validation samples. To compute evaluation scores, we consider only the source speech and target texts, the complete evaluation pipeline is described in section 4.3.

The unit-to-speech (U2S) model is separately trained from S2UT model. To train the U2S system for English language, we combine the LJSpeech dataset [21] and the ESD[22] dataset. The LJSpeech dataset contains 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books, with a total duration of approximately 24 hours. ESD is a multilingual emotional database, consisting of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers (10F, 10M). In this study, we only consider the English part.

4.1. Baseline

To assess the effectiveness of our proposed approach, we build a Baseline model which is composed of a speech-to-unit translation (S2UT) module and a unit-to-speech (U2S) module. We conduct an analysis by systematically excluding the emotion encoder and pitch predictor from the U2S module. This enables us to quantify the impact and measure the benefits of their inclusion in the overall system.

4.2. TTS

In addition to the Baseline and our S2ST model, we also incorporated an English text-to-speech (TTS) model [23] into our subjective evaluation. The TTS model was trained on the identical dataset utilized for training the U2S module. Its inclusion serves to assess the overall quality of the synthesized speech generated by the TTS model in comparison to our proposed system and to evaluate the effectiveness of expressivity transfer achieved by our proposed system in contrast to the TTS model.

4.3. Evaluation

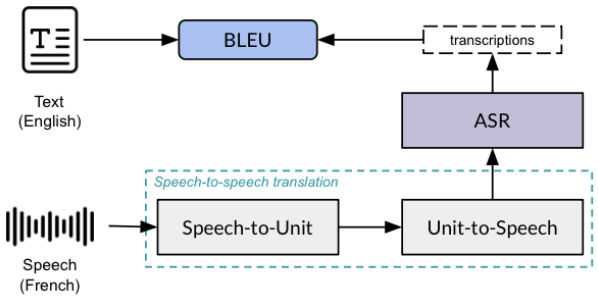

Fig. 2. An illustration of the evaluation pipeline used for speech-to-speech translation.

Recent work in speech-to-speech translation suggests to evaluate translation quality using the BLEU score. We start by using an ASR model to compute the transcriptions of the generated speech. In order to obtain comparable results, we use the same open source ASR model as in [17]. Then, we compute BLEU score of the ASR decoded text with respect to the reference translations. We acknowledge that the ASR BLEU score may not be a perfect metric for assessing data quality, as it will be unavoidably influenced by the performance of ASR models. The complete evaluation pipeline of speech-to-speech translation is illustrated in Figure 2.

Table 1. BLEU scores on EPST and FLEURS test sets by S2ST models with different settings

Model	BLEU	
	EPST	FLEURS
Synthetic target	82.6	82.7
Baseline	17.0	15.7
S2ST	17.3	15.9
Baseline *multitask*	16.7	14.0
S2ST *multitask*	17.0	14.2
From the literature:		
SpeechMatrix	20.7	9.8

In addition to measuring the translation quality via an objective metric, we conduct human listening tests to assess perceptual responses of expressivity transfer from recordings generated by our S2ST model. We asked 33 people to evaluate two sets of tasks online. A detailed description of the tasks was provided to all evaluators, who had unlimited time to evaluate audio stimuli. Each task was organized similarly, consisting of one pre-trial (excluded from this analysis) and four trials. Each trial contained three synthesized speech recordings produced by Baseline, TTS, and our S2ST framework. After listening to each recording, evaluators provided an opinion score on a scale of 1 to 5, where 1 is 'Poor' and 5 is 'Excellent'. The first task was a Mean Opinion Score (MOS) where evaluators judged the quality of the synthesized speech. The second task was a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), where evaluators listened to a reference (natural, spoken French) and then judged the expressiveness of the English-translated synthesized speech.

5. RESULTS

We first evaluate translation quality of the Baseline and the S2ST model using BLEU score (Section 5.1). Next, we conduct a subjective evaluation in terms of audio-quality (MOS) along with expressivity transfer (MUSHRA) and compare the proposed method against the Baseline and TTS (Section 5.2).

5.1. Speech-to-speech translation

We investigate the training of a speech-to-speech translation system using both single and multitask learning approaches. Table 1 summarizes performance of S2ST models on both EPST and FLEURS test sets. We include the results from SpeechMatrix [17] as references as the exact same ASR models is used for evaluation. Additionally, we present the BLEU scores calculated for the synthetic target speech to show the impact of ASR errors on the evaluation metric.

First, we compare the proposed S2ST model to the Baseline. We can see that our S2ST model outperforms the Baseline by 0.3 BLEU on EPST and by 0.2 BLEU on FLEURS, indicating that our approach performs similar or slightly better in terms of translation performance. We also note that SpeechMatrix achieves an improvement of 3.4 BLEU over the proposed S2ST model on EPST, however, on FLEURS our approach outperforms SpeechMatrix by 6.1 BLEU leading to an average improvement of 1.3 BLEU. The gap of performance on the FLEURS test set can be attributed in part to the fact that we use an encoder pre-trained on 7000 hours of speech coverings multiples domains compared to SpeechMatrix encoder trained only on European Parliament recording.

Secondly, we explore multitask learning by incorporate an auxiliary task to the Baseline and S2ST model. In our experimental setup, we observe a decline in performance for both the Baseline and the S2ST model when employing multitask learning. Specifically, the S2ST model yields a performance of (17 vs. 17.3) on EPST and (14.2 vs. 15.9) on FLEURS. This suggests that our encoder does not provide significant benefits to the auxiliary task. Nonetheless, our approach still outperforms the Baseline system for both setups, indicating the effectiveness of our proposed approach.

5.2. Subjective Evaluation

Separate linear mixed effects models were used to evaluate MOS and MUSHRA task responses. Using the R-package $lme4$, opinion responses were entered as response variables. Synthesized speech system (3-levels) and speaker sex (2-levels) were entered as fixed factors and participant was entered as a random factor. Chi-squared ($\chi^2_{d,N}$) tests were used to report p-values ($Anova$ from the car R-Package) with d degrees of freedom and N samples, i.e., there were N = 486 responses, d = 2 speech systems, and d = 1 speaker sexes. Main effects were reported for task, response, and their interactions with speaker. Estimated marginal means ($emmeans$) were used to conduct pairwise comparisons, where $X \pm Y$ represent mean and standard error, respectively.

The results of the MOS task revealed significant main effects on system $\chi^2_{2,486}$ = 284.17 and speaker sex $\chi^2_{1,486}$ = 11.25, as well as their interaction $\chi^2_{2,486}$ = 18.66, $p < 0.001$. Pairwise comparisons showed that the quality of recordings generated by the Baseline system (2.07 $\pm$ 0.1) had significantly lower opinion scores in comparison to those generated by TTS (3.45 $\pm$ 0.1) and our S2ST model systems (3.56 $\pm$ 0.1), $p < 0.001$. In comparison to female speech recordings (2.89 $\pm$ 0.09), male speech recordings (3.17 $\pm$ 0.09) had significantly increased scores, $p < 0.001$, however, these effects were localized to the Baseline system (Figure 3 Left-Middle).

MUSHRA task results showed a significant main effect on system $\chi^2_{2,453}$ = 14.27, $p < 0.001$, but not on speaker sex, $p >$ 0.05. Pairwise comparisons showed that the expressiveness of recordings generated by our S2ST model (3.08 $\pm$ 0.11) had significantly increased higher opinion scores in comparison to those generated by Baseline (2.66 $\pm$ 0.11) and TTS systems (2.69 $\pm$ 0.11), $p < 0.01$ (Fig. 3-Right).

To better understand the MUSHRA task results, OpenSmile was used to extract 88-acoustic features (eGeMAPS [24]) that were entered in a forward SLDA with Wilks'

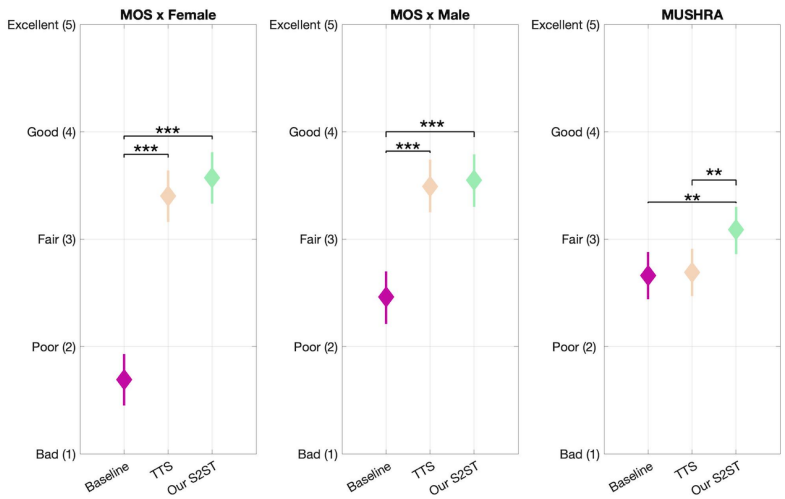

Fig. 3. MOS (Left-Middle) and MUSHRA (Right) task results. Diamonds and vertical lines represent mean and critical intervals. {**, ***} represent $p < \{0.01, 0.001\}$.

Lambda criterion (R function $greedy.wl$ in the $klaR$ R-package). In order to identify which acoustic features distinguished our S2ST model from Baseline and TTS systems, a forward SLDA method was preferred, as it starts from the null hypothesis and incrementally adds new variables with the highest discriminant power based on Wilks' Lambda value until $p > 0.01$. Based on the six acoustic features selected[2] from the SLDA, standardized euclidean distances were computed between the reference (French) and synthesized (English) speech recordings produced by Baseline, TTS, and the proposed S2ST model. One-sided ANOVA results revealed a significant effect of system on euclidean distances between reference and synthesized speech $F_{2,87} = 4.11$, $p < 0.05$. Pairwise comparisons showed our S2ST model (2.75 $\pm$ 0.25) had significantly smaller distances in comparison to the TTS system (3.75 $\pm$ 0.25), however, no differences from the Baseline (3.16 $\pm$ 0.25). Finally, Pearson correlation procedures showed a significant relationship between mean opinion and euclidean distance between the reference and our S2ST model (ρ = -0.39, $p < 0.05$).

There are several takeaways from our subjective evaluations. First our S2ST framework produced speech recording that were perceived to have higher quality in comparison to those produced by the Baseline system. Next it outperformed both Baseline and TTS systems in terms of producing recordings that conveyed speaker expressivity. The euclidean distances of a select set of acoustic features (6) extracted from reference and our S2ST model speech recordings were found to be significantly smaller in comparison to TTS system and negatively correlated to opinion scores.

[2]The following acoustic features were selected (with Wilks' lambda and F-stat values): slopeUV500_1500_sma3nz_amean (λ: 0.22; F: 80.26), slopeV0_500_sma3nz_amean (λ: 0.06; F: 56.2), F0semitoneFrom27_5Hz_sma3nz_stddevRisingSlope (λ: 0.03; F: 17.16), F1bandwidth_sma3nz_amean (λ: 0.02; F: 14.98), logRelF0_H1_H2_sma3nz_amean (λ: 0.01; F: 17.47), and mfcc4V_sma3nz_stddevNorm (λ: 0.01; F: 8.88).

6. CONCLUSIONS

In this paper, we have addressed the crucial challenge of preserving expressivity in speech-to-speech translation systems. Our proposed approach leverages multilingual emotion embeddings, resulting in significant advancements in retaining the nuances of expressiveness during textless translation. The experimental results have demonstrated the superior expressivity transfer achieved by our method compared to state-of-the-art systems, highlighting its effectiveness.

Moreover, our speech-to-speech translation framework has produced speech recordings that were perceived by humans to have higher quality in terms of conveying speaker expressivity, surpassing both our speech-to-speech Baseline and text-to-speech systems. Importantly, we have maintained the translation quality at a level similar to that of state-of-the-art textless speech-to-speech translation systems.

Looking ahead, future research directions involve exploring the incorporation of additional paralinguistic information, optimizing the generation of speech discrete units for this task, and expanding the approach to other language pairs, particularly unwritten ones.

7. ACKNOWLEDGMENTS

This work received funding from the European SELMA project (grant N°957017).

8. REFERENCES

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and Z. Puming, "Janus-iii: speech-to-speech translation in multiple languages," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, vol. 1, pp. 99–102.

[2] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The atr multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[3] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[4] Y. Jia, R. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Interspeech 2019*, 2019.

[5] A. Lee, P. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W. Hsu, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 3327–3339.

[6] A. Lee, H. Gong, P. A. Duquenne, H. Schwenk, P. J. Chen, C Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W. N. Hsu, "Textless speech-to-speech translation on real data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, Association for Computational Linguistics.

[7] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W. N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech 2021*. 2021, ISCA.

[8] J. Duret, Y. Estève, and T. Parcollet, "Learning multilingual expressive speech representation for prosody prediction without parallel data," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[9] K. Lakhotia, E. Kharitonov, W. Hsu, Y. Adi, A. Polyak, B. Bolte, T. Nguyen, J. Copet, A. Baevski, and A. Mohamed, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[10] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. Nguyen, M. Rivière, W. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11200–11214.

[11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020.

[12] S. Popuri, P. Chen, and C. et al. Wang, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[13] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv*, 2021.

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv*, 2020.

[15] K. Kasi and S. Zahorian, "Yet another algorithm for pitch tracking," in *IEEE International Conference on Acoustics Speech and Signal Processing*. 2002, IEEE.

[16] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.

[17] P. Duquenne, H. Gong, and N. et al. Dong, "Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations," *arXiv preprint arXiv:2211.04508*, 2022.

[18] J. Iranzo-Sánchez, J. Silvestre-Cerda, and J. et al. Jorge, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8229–8233.

[19] A. Conneau, M. Ma, and S. et al. Khanuja, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[20] N. Goyal, C. Gao, and V. et al. Chaudhary, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022.

[21] I. Keith and J. Linda, "The lj speech dataset," 2017.

[22] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," 2022.

[23] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[24] F Eyben, M Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010.