Toward transfer of acoustic cues of emphasis across languages

Andreas Tsiartas, Panayiotis G. Georgiou and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab,
Ming Hsieh Department of Electrical Engineering,
University of Southern California, Los Angeles, CA 90089

tsiartas@usc.edu, <georgiou,shri>@sipi.usc.edu

Abstract

Speech-to-speech (S2S) translation has been of increased interest in the last few years with the research focused mainly on lexical aspects. It has however been widely acknowledged that incorporating other rich information such as expressive prosody contained in speech can enhance the cross-lingual communication experience. Motivated by recent empirical findings showing a positive relation between the transfer of emphasis and the quality of the audio translation, we propose a computational method to derive a set of acoustic cues that can be used in transferring emphasis for the English-Spanish language pair. In particular, we present an iterative algorithm that aims to discover the set of acoustic cue pairs in the two languages that maximize the accurate transfer of emphasis. We find that the relevant acoustic cues can be constructed from a diverse set of features including word/phrase level statistics of spectral, intensity and prosodic cues and can model the acoustic information related to emphasized and neutral words/phrases for the English-Spanish language pair. These features can in turn enable data-driven transformations from source to target language that preserve such rich prosodic information. We demonstrate the efficacy of this approach through experiments on a specially constructed corpus of 1800 English-Spanish words/phrases.

Index Terms: Speech-to-speech translation, paralinguistic cues representation

1. Introduction

Speech-to-speech (S2S) translation's ultimate goal is to allow spoken human communication across different languages, dialects and cultures. S2S is becoming more desirable due to increasingly multicultural societies, people's increased travel, and due to widely available Internet-connected devices such as smart-phones. The need is also evident in improving healthcare delivery among patients and doctors that do not speak the same language [1]. This need has attracted research and industry towards the creation of a robust and accurate S2S translation system.

A variety of S2S systems have been proposed in the literature [2, 3, 4]. A typical speech-to-speech (S2S) system is composed of an automatic speech recognizer (ASR) which converts the input into words, the words are translated using a statistical machine translator (SMT) and, finally, a Text-To-Speech (TTS) system is used to compose the target signal. In such pipelined S2S approach, one can isolate and work on subsystems independently. However, S2S translation is beyond this pipelined S2S approach. Recent work [5] has shown that additional paralinguistic cues such as emphasis can be also useful for S2S translation.

There is limited systems-side work in bringing paralinguistic aspects into S2S translation. However, there is early research into exploiting paralinguistic cues in the S2S framework. Parlikar *et al.* [6] have used phoneme mappings as acoustic units to adapt the TTS output signal from the input language and shown benefits on the TTS side. On the feature side, power and duration have been used in [7] to translate emphasized digits and wherein the prediction of emphasis and root mean squared error rate (RMSE) have been used as an evaluation metric. Aguero *et al.* [8] have used an unsupervised approach in learning prosodic mappings and showed TTS output benefits in terms of mean opinion score. Rangarajan *et al.* [9] used dialog acts and prosodic cues obtained from the input speech signal within the SMT component and have shown translation benefits in terms of BLEU score [10].

While such approaches can offer useful information to various aspects of S2S components, a computational approach to learn paralinguistic representations can be very important for all components and S2S translation in the same way phonemes and words are useful for ASR. In contrast to existing work that focuses on the entire S2S system to show improvement in terms of different aspects of S2S translation, in this paper, we focus on deriving acoustic representations that maximize the direct information transfer across languages. We present a data-driven supervised approach that learns acoustic mappings by discretizing the acoustic space (modeled by diverse speech features such as MFCCs, pitch etc.) with the K-Means algorithm. The code mappings are evaluated using the mutual information between the bilingual discrete representations and the presence of paralinguisticly salient. In addition, the bilingual acoustic representations are evaluated by conditional entropy to measure the uncertainty of the mappings.

Specifically, in this paper, we show the efficacy of the approach by creating a representation for prosodic information transfered and focus on deriving the most informative acoustic representations. The representation is created from acoustic feature vectors discretized and evaluated using mutual information shared between the representation and the emphasis transfer. The representation is learned from a quadruplets of parallel utterances spoken in neutral (flat tone) English, neutral Spanish, and English, Spanish with appropriate emphasis. In addition, we further attempt to jointly maximize the information transfered and the predictability of the encoding using conditional entropy as a measure.

This paper is structured as follows. In section 2, we explain how the 4-way bilingual data have been collected. In section 3, we describe the acoustic measures used to create the acoustic representation. In section 4, we give a brief description of the word/phrase level features used to model the acoustic space.

Section 5 describes the approach used to map the acoustic space to the acoustic representation. In section 6, we describe the experimental setup and section 7 discusses the results of this work. Finally, in section 8, we summarize the findings of this work and provide some future directions.

2. Data-Driven Learning

To collect data suitable for directly learning emphasis transfer representations for the English-Spanish bilingual pair, we recruited two bilingual actors, one male and one female. We obtained a random utterance set from the IEMOCAP [11] database and translated all English utterances to Spanish.

The utterances were tagged with words to be emphasized. The corresponding word/phrase on the translated Spanish side has been marked as well. The actors spoke the utterance in both languages with emphasis and neutral resulting into a quadruplet. We recorded 450 such quadruplets resulting in 1800 utterances. Next, we extracted the words/phrases that are emphasized with their neutral counterparts in both languages resulting into 1800 words/phrases. The set has been split into half for training and half testing.

3. Acoustic Representation

In this section, we propose a representation for the acoustic cues transfered in cross-lingual spoken translation. To create this representation, we propose a mapping from the continuous acoustic space of speech to a discrete set of acoustic units.

Hence, say we have two words/phrases spoken in two languages L_1 and L_2. Let X_{L_1} and X_{L_2} be signal representations of the words/phrases, we define two mappings independently for the two languages to yield the corresponding discretized (quantized) vectors as follows:

$$X_{L_1} \rightarrow A_{L_1}$$

and

$$X_{L_2} \rightarrow A_{L_2}$$

The signal representation X_{L_i} is composed of a set of features, for example, transformations of MFCC, pitch and other spectral and prosodic features. The mapping defines a discretization of such features which denoted as A_{L_i}. To construct such a mapping, we use K-means clustering [12] to map the continuous space of acoustic cues to a finite discrete set of acoustic units.

3.1. Transfer of acoustic cues

With the aforementioned representation, we need a way to measure how well cues are transfered by the particular representation. Each feature vector and mapping to acoustic units can create a representation in which some mappings can model the "language" of acoustic cues transfered. Thus, we propose to use an information theoretic approach to evaluate each representation created by different feature vectors and different mappings to acoustic units. Hence, given a perceptual acoustic transfer Y, for example emphasis transfer, we need to find a representation $A = (A_{L_1}, A_{L_2})$ such that their mutual information is maximized:

$$I(A, Y) = \sum_{a \in \mathcal{A}, y \in \mathcal{Y}} P(a, y) log \frac{P(a, y)}{P(a)P(y)}$$

where P defines the probability measure.

3.2. Conditional entropy for minimizing code uncertainty

While a specific representation can model the information shared between a language pair for analysis purposes of the acoustic cues transfered, it might have high uncertainty in the translation process. Thus, it is useful to have a measure to model the coding mapping uncertainty. For this reason, we propose to use a soft metric to evaluate how well the acoustic translation representation can be predicted using conditional entropy which can be written as:

$$H(A_{L_1}|A_{L_2}) = \sum_{a_{L_1} \in \mathcal{A}_{L_1}, a_{L_2} \in \mathcal{A}_{L_2}} P(a_{L_1}, a_{L_2}) log \frac{P(a_{L_2})}{P(a_{L_1}, a_{L_2})}$$

Using the conditional entropy metric, we can evaluate the ambiguity of the coding scheme.

4. Acoustic features

We considered a variety of acoustic feature vectors (X_{L_1},X_{L_2}) to represent different aspects of the speech spectrum and prosodic cues. All features are defined at the word/phrase level.

4.1. Mean power

The first feature we used in our representation is mean power to model the transfer of emphasis. Since power has been used widely in a variety of settings for modeling emphasis, we use it to produce a baseline representation.

4.2. Additional features

In addition to mean power, we have used word/phrase duration and various word/phrase level statistics of features which include MFCC, voicing pitch, etc. Statistics used include quantiles, mean, max, min, etc. In total we have extracted 6126 word/phrase level features for each word/phrase. The feature set has been extracted using OpenSMILE [13] as used in [14].

5. Acoustic unit estimation approach

In this section, we describe the approach used to create the acoustic units. A basic layout of the approach is shown in Fig. 1. The approach is iterative and the dimensionality of the feature space is increased progressively and in a greedy manner.

The algorithm is initialized with an empty feature vector. In step one, we add a feature to the feature vector $\mathcal{F}$. In the second step, the K-Means [12] algorithm is run independently for languages $L1$ and $L2$ and the encoding (A_{L_1} and A_{L_2}) is created for each language. Thirdly, the coding is evaluated using the MI and conditional entropy metric as defined in Sec. 3. If the metric considered improves, the feature replaces the last feature added in $\mathcal{F}$. If all features described in 4 are exhausted, we increase the dimensionality of $\mathcal{F}$ and go to step one.

6. Experimental setup

For experimenting with the emphasis transfer problem, we run the algorithm described in Sec. 5 in different setups. First, we run the algorithm by maximizing the mutual information. Then, we run the algorithm by maximizing the mutual information between the cross-lingual acoustic representations and the emphasis transfer at each step and at the same time minimizing the entropy so that we include as much information about the paralinguistic cue transfer but also find a representation that will

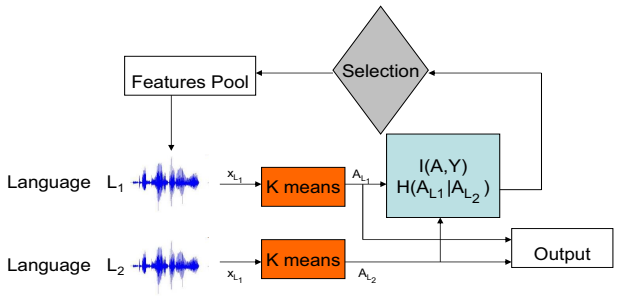

Figure 1: The iterative approach used to find the best acoustic representation for the acoustic cues transfered.

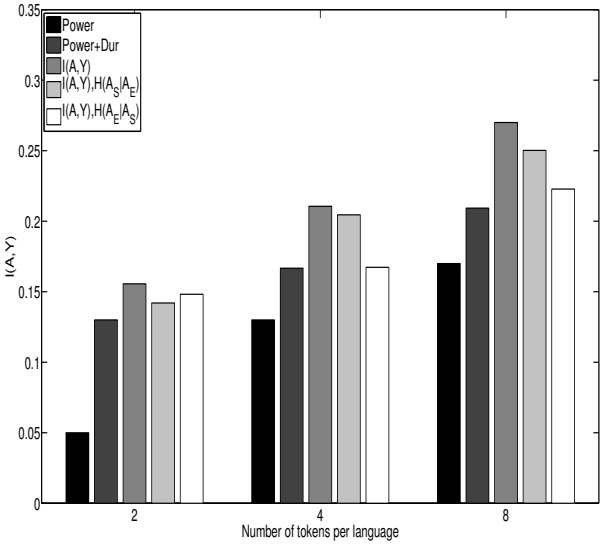

Figure 2: This figure shows the mutual information $I(A, Y)$ of the acoustic representations for emphasis transfer for different approaches and different number of tokens.

minimize the coding prediction error. In addition, as described in Sec. 4, we used the mean power and duration of the signal to create a representation and form a baseline to evaluate the efficacy of our approach. To perform this optimization, we split the data set into two parts one for training and one for testing with half of the data in each set. The optimization has been run on the training set and we report the results on the testing set. Since the acoustic representations are created on the training set, we assign to the testing feature vector the closest code as defined by its cluster center.

Finally, we repeated the experiments for coding schemes with vocabulary size of 2, 4 and 8 codes in each language. For computational purposes, we run the experiments until at maximum 10 features are added or stopped if no more features can be added to improve the metrics considered.

7. Results and Discussion

In this section, we analyze the results of the computational approach to find appropriate representations for the English-Spanish pair. Fig. 2 shows the value of mutual information (MI) between the acoustic representations in English and Spanish and the transfer of emphasis. Results show that the amount of information transfered increases when increasing the number of tokens the acoustic representation is composed. Adding the duration to the power baseline (Power+Dur) can increase the information transfered in the English-Spanish bilingual pair.

Furthermore, when we maximize the mutual information ($I(A, Y)$) the approach can identify acoustic representations that yield up to 0.07 MI measure higher than power and duration together depending on the number of tokens considered which in-turn implies that more information is transfered. Using the Wilcox rank test and breaking the test set into 45 subsets, we find that the results are significant at p-values less than 10^{-8} for both the comparisons.

In applications such as speech translation, it is important not only to ensure that the representations ensure maximal information transfer, but yield as minimal ambiguity as possible to enable correct translation with low uncertainty. For this reason, we repeated our experiments by maximizing the mutual information ($I(A, Y)$) and minimizing the entropy ($H(A_{L_S}|A_{L_E})$ and vice versa for Spanish→English) at each step of the algorithm.

Results in Fig. 2 show that such joint optimization yields less information carried in the bilingual English-Spanish pair than optimizing only on mutual information ($I(A, Y)$) by up to 0.05 depending on the number of tokens considered but still more information than the baselines of up to 0.1 in terms of mutual information.

In addition, while adding duration to the baseline increased the mutual information, in three cases it increased the ambiguity of the coding scheme only when discretizing into two representations. Also, as shown in Fig. 3 the computational approach to create acoustic representations yielded codes with much lower ambiguity measured in terms of conditional entropy. In particular, this dual metric can lower the conditional entropy by up to 0.7 points depending on the number of tokens considered. The improvements in conditional entropy are consistent for both sides of the mapping of the acoustic information for all numbers of tokens considered and results are significant at p-values less than 10^{-8} for all cases.

While optimizing only on MI, the conditional entropy remains very close to the baseline but with much more cross-lingual information carried in the representation.

8. Conclusion

In this work, we presented a computational approach to construct a cross-lingua representation for acoustic cues transfer and, in particular for the emphasis transfer. We have presented a mapping from the acoustic feature space to a discrete set of units using an iterative procedure in which at each step the mutual information is maximized. This method can potentially lead to an approach to learn cross-lingual information across speech-to-speech (S2S) components that can be used beyond the pipelined architecture of S2S by exploiting a diverse set of features.

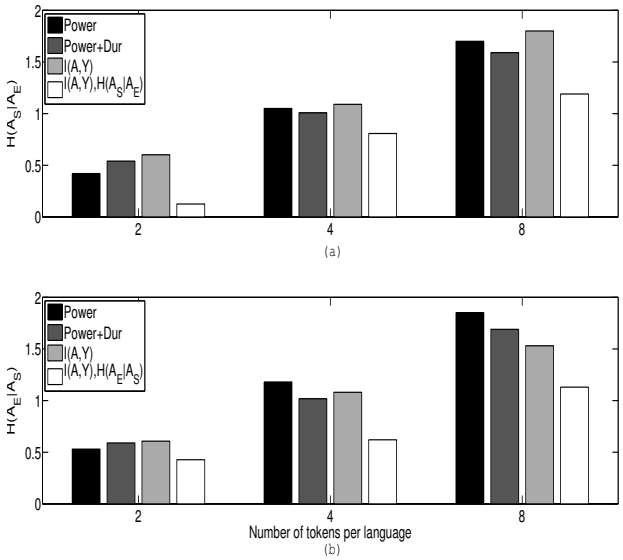

Figure 3: This figure shows the conditional entropy for the English to Spanish (a) and Spanish to English (b) translation of the acoustic representations with different approaches for different number of tokens.

Furthermore, for applications such as speech translation that require the resulting acoustic units to have low uncertainty for the prediction while simultaneously transferring as much information as possible, we added another condition to the algorithm to jointly maximize mutual information and minimize the conditional entropy of the prediction. Our results indicate that for applications in which the information transfered is important we can achieve MI up to three times higher than the baseline considered.

In addition, for applications requiring not only the maximum amount of information transfered but also low ambiguity of the coding scheme, the joint maximization of mutual information and conditional entropy yielded reductions in terms of conditional entropy of up to 3.5 times for the English Spanish bilingual translation.

For future work, we intend to collect and evaluate our approach on more speakers. In addition, we want to explore more features sets that can be used in the approach and also improve the approach with different optimization techniques to yield higher mutual information (MI) and lower conditional entropy as a measure of the coding scheme uncertainty. Also, additional metrics can be useful for extracting different cross-lingual information useful in different S2S components.

9. References

[1] B. D. Smedley, A. Stith, and A. Nelson, "Unequal treatment: Confronting racial and ethnic disparities in health care.," *Institute of Medicine Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care.*, 2003.

[2] Zheng J., Mandal A., Lei X., Frandsen M., Ayan NF, Vergyri D., Wang W., Akbacak M., and Precoda K., "Implementing SRI's Pashto speech-to-speech translation system on a smart phone," in *Spoken Language Technology Workshop (SLT), 2010 IEEE*. IEEE, 2010, pp. 133–138.

[3] Y. Gao, Gu L., Zhou B., Sarikaya R., Afify M., Kuo H.K., Zhu W., Deng Y., Prosser C., Zhang W., et al., "IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator," in *Proceedings of the Workshop on Medical Speech Translation*. Association for Computational Linguistics, 2006, pp. 53–56.

[4] Prasad R., Krstovski K., Choi F., Saleem S., Natarajan P., Decerbo M., and Stallard D., "Real-time speech-to-speech translation for Pdas," in *Portable Information Devices, 2007. PORTABLE07. IEEE International Conference on*. IEEE, 2007, pp. 1–5.

[5] Tsiartas A., Georgiou P., and Narayanan S., "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *Proc. IEEE ICASSP*. IEEE, 2013.

[6] A. Parlikar, A. Black, and S. Vogel, "Improving speech synthesis of machine translation output," in *INTERSPEECH*, September 2010, pp. 194–197.

[7] Takatomo K., Sakriani S., Shinnosuke T., Graham N., Tomoki T., and Satoshi N., "A method for translation of paralinguistic information," *Proceedings IWSLT 2012*, 2012.

[8] P. D. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *ICASSP*, Toulouse, France, May 2006.

[9] V. Rangarajan, S. Bangalore, and S. Narayanan, "Enriching machine-mediated speech-to-speech translation using contextual information.," *Computer Speech and Language*, 2011.

[10] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics (ACL)*, 2002, pp. 311–318.

[11] Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J.N., Lee S., and Narayanan S., "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[12] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2000.

[13] Florian E., Martin W., and Björn S., "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[14] Björn S., Stefan S., Anton B., Elmar N., Alessandro V., Felix B., Rob V., Felix W., Florian E., Tobias B.and Mohammadi G., and Weiss B., "The interspeech 2012 speaker trait challenge," *Interspeech, Portland, Oregon*, 2012.