A STUDY ON THE EFFECT OF PROSODIC EMPHASIS TRANSFER ON OVERALL SPEECH TRANSLATION QUALITY

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

Despite the increasing interest in Speech-to-speech (S2S) translation, research and development has focused almost exclusively on the lexical aspects of translation. The importance of transferring prosodic and other paralinguistic information through S2S devices and evaluating its impact on the translation quality are yet to be well established. The novelty in this work is a large scale human evaluation study to test the hypothesis that cross-lingual prosodic emphasis transfer is directly related to the perceived quality of speech translation. This hypothesis is validated at the 0.53-0.54 correlation level on the data sets considered with results significant at p-value=0.01. The second contribution of this work is an evaluation methodology based on crowd sourcing using English-Spanish language bilingual data from two distinct domains and evaluated with over 200 bilingual speakers. We also present lessons learned on this type of S2S subjective experiments when using crowd sourcing.

Index Terms— Signal to speech translation, paralinguistic emphasis translation

1. INTRODUCTION

The goal of Speech-to-speech (S2S) translation is to allow spoken human interactions across different languages and support communication between people with limited or no knowledge of a certain spoken language. Such need is felt widely in today's increasingly multilingual multicultural world, such as in improving delivery of health-care to patients that do not share the same language [1]. Also due to the rapid expansion of tourism, Internet and smart-phones, S2S translation has attracted researchers attention during the last decade for building and using S2S translation applications that are portable and personal [2, 3, 4].

A typical S2S system has a pipelined architecture [5] in which an automatic speech recognizer (ASR) receives the speech signal and converts it into a sequence of words. Then, the sequence of words is translated with the statistical machine translator (SMT) into the target language. Finally, the words in the target language are synthesized using a Text-to-Speech (TTS) system. This pipelined architecture has its advantages in the sense that each component of this S2S pipeline can be isolated and researched independently. However, it has limitations when additional source language information needs to be exploited and for which the individual components are not designed to model.

1.1. Relation to prior work

Only limited work has been done in incorporating information that is not supported by the aforementioned components of the typical S2S pipelined architecture. In some works, additional information extracted from the speech signal has been used within the S2S components individually. For example, Parlikar *et al.* [6] have adapted the TTS output signal using phoneme mappings from the input language and have shown TTS improvement. Aguero *et al.* [7] used an unsupervised method to learn prosodic mappings trained on bilingual read sentences which are then used to enhance the TTS output and have shown benefits in terms of mean opinion score. Rangarajan *et al.* [8] have added dialog acts and prosodic features obtained from the source signal in the SMT component and have shown translation improvements in terms of BLEU score [9].

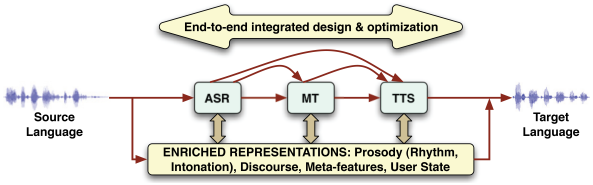


Fig. 1. A system architecture that can exploit speech information beyond the pipelined architecture used in speech-to-speech systems.

The importance of paralinguistic information in monolingual human communication has been widely documented [10, 11, 12]. The premise of our work is that such paralinguistic information is important in cross-lingual communication settings, and S2S systems should possess such capability. What is however unclear is what aspects of the multifaceted rich information in the source speech would be beneficial for inclusion in the cross-lingual transfer. Toward that end, in this paper we describe a method to systematically explore and evaluate the role and importance of specific aspects of paralinguistic information in S2S translation. This can be viewed as a design step even before an actual technology system is created.

We perform perceptual evaluation experiments using a crowd sourcing approach widely used in various experimental settings in the past [13, 14, 15, 16]. In particular, we describe a case study aimed at investigating whether the transfer of emphasis of a word or a phrase in the source language to the target language is related to the perceived translation quality. To carry out our experiments, we use bilingual utterances obtained from dubbed movies and doctor-patient-interpreter interactions, with subjective experiments on Amazon Turk[1] to test our hypothesis. In this paper, we focus on the English-Spanish language pair. We find that the transfer of emphasis significantly correlates with the perceived translation quality for this language pair. In addition, we describe our experience in evaluating this type of S2S experiments using crowd sourcing.

This paper is structured as follows. In section 2, we describe the

[1] http://www.mturk.com

data collected and used in this work. In section 3, we elaborate on our hypothesis. Section 4 describes the survey used to conduct the perceptual experiments. Section 5 presents the experimental setup on Amazon Turk. In section 6, we discuss the results of this work and, finally, in section 7, we summarize the findings of this work and provide some future directions.

2. DATA COLLECTION

In this section, we describe the data sets collected for testing our hypothesis through human evaluation experiments. We focus on two different data sets.

2.1. S2S data set

The first data set was collected by SAIL[2] as a part of a medical domain S2S translation project called Speech-links. This data set involves interactions between an English speaking doctor, a Spanish speaking patient and a bilingual interpreter that facilitates this interaction by translating from English to Spanish and vice versa. The doctors are students from USC's Keck School of Medicine and the patients are standardized patients. This method originally was proposed in [17]. The interpreters are professional English-Spanish interpreters trained to facilitate medical interactions in California's hospitals. The recordings took place in a typical room setting with little background noise coming from air-conditioning etc. Each session lasted up to half an hour. There are a total of six doctors, six interpreters and six patients and at the end of each doctor-patient-interpreter setting, and participants are permuted to ensure variety in the pairs involved in the interaction. The interactions are highly realistic and spontaneous and the interpreters were unconstrained in their task with minimal instructions that they should attempt to minimize overlap.

For the purpose of these experiments, we hired English-Spanish bilingual speakers to randomly pick 50 bilingual utterances from 10 different sessions resulting in a total set of 500 utterance pairs. We also asked the bilingual speakers to manually transcribe and match them in bilingual pairs, for example, two utterances are put together if they are bilingual translations of each other in the interaction. From now on, we will refer to this data set as S2SData set.

2.2. Movies data set

The second data set we experimented with comprises bilingual utterances that are from dubbed movies. This type of utterance pairs has more or less the same duration as the source utterance due to constraints of the visual channel. Often the translations are made in a way to lip-sync the words spoken in the source language. Dubbed movies are processed off-line and dubbed by professional interpreters, with the possibility of being recorded multiple times and, also, if possible, lip synced to match the video both in timings and visuals. In this sense, dubbed movies data differ from the S2SData set.

To obtain a set of high-quality bilingual utterances, we segmented the data using the approach described in [18]. Then, we processed the bilingual utterances and selected the clean bilingual pairs that do not contain background noise. From 15 dubbed movies, we randomly selected 781 clean bilingual utterances, ensuring that the pairs were conceptually translations of each other, and transcribed them manually in both languages. From now on, we will refer to this data set as Movies data set.

[2]http://sail.usc.edu

3. HYPOTHESIS: TRANSFER OF PROSODIC EMPHASIS

Our goal is to examine the hypothesis whether translation quality is affected by the quality of transfer of paralinguistic cues. In particular, we focus on the transfer of emphasis. In signal processing terms, emphasis/stress is defined as the perceived loudness of a word/phrase. Intuitively, if we want to emphasize a word/phrase, or a concept, we stress specific words/phrases of the utterance. By stressing the word/phrase, we may change the meaning conveyed by the utterance and, thus, such cues have to be taken into consideration in the translation. For example, an emphasized word might be important in the context of the dialog and the annotator might need to pay special attention to that word/phrase.

Our main premise is that this paralinguistic cue is important both in terms of production (interpreters transfer this information) and in terms of perception (annotators perceive this information). We perform perceptual evaluation experiments to test this hypothesis for the English-Spanish language pair.

4. PERCEPTUAL EVALUATION EXPERIMENTS

To perform the perceptual experiments and test our hypothesis, we used the data described in section 2 and created the survey shown in Fig. 2. At each instance, we provided one bilingual Spanish-English utterance pair to the annotators and asked them to rate the quality of the translation (on a scale 1-5, question 1) and how well the emphasis of the English audio is preserved in the Spanish audio (on a scale 1-5, question 3). In addition, the annotators were asked to give their confidence in rating the emphasis preservation (on a scale 1-5, question 4) and, whether, they perceive any words/phrases that are emphasized in the English audio (yes/no answer, question 2).

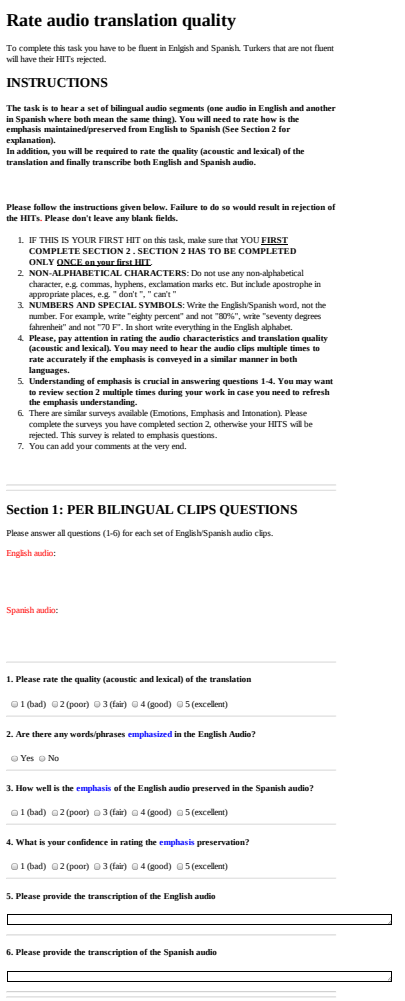
To examine our hypothesis, we tested the relation between the results on the quality of translation (question 1) with the ratings of emphasis preservation (question 3). Confidence ratings (question 4) were used to examine the hypothesis above for confident annotations. In addition, for quality testing purposes, we asked the annotators to transcribe each utterance (questions 5-6), thus, ensuring they paid attention to each audio signal. The survey given to the annotators is shown in Fig. 2[3].

5. EXPERIMENTAL SETUP

For the perceptual experiments, we employed crowd sourcing through Amazon Turk. Using the survey described in section 4, we requested that each annotator participating in the survey to be an English-Spanish bilingual speaker. Before filling the survey, each annotator was mandated to go through training. Annotators were presented with samples of speech containing emphasized words/phrases and samples with no emphasized words/phrases so that we were sure it was clear to the annotators what is the definition of emphasis. In addition, they were presented examples in which emphasis was transfered and other cases that emphasis was not transfered. Their attention to training was ensured through monitoring of the accuracy in transcription of each utterance that they had to transcribe. At this point, we accepted annotators that passed the training section without transcription errors.

Annotators who cleared the training phase had to answer the four questions explained in section 4 and to transcribe the utterances in both languages (Fig. 2, questions 5-6). To ensure the quality of the tagged data, we rejected annotators having Word Error Rate (WER) [19] greater than 25.0%.

[3]Note that a one-time demographic survey and training session was given to each annotator



Rate audio translation quality

To complete this task you have to be fluent in Enlgish and Spanish. Turkers that are not fluent will have their HITs rejected.

INSTRUCTIONS

The task is to hear a set of bilingual audio segments (one audio in English and another in Spanish where both mean the same thing). You will need to rate how is the emphasis maintained/preserved from English to Spanish (See Section 2 for explanation).
In addition, you will be required to rate the quality (acoustic and lexical) of the translation and finally transcribe both English and Spanish audio.

Please follow the instructions given below. Failure to do so would result in rejection of the HITs. Please don't leave any blank fields.

1. IF THIS IS YOUR FIRST HIT on this task, make sure that YOU **FIRST COMPLETE SECTION 2 . SECTION 2 HAS TO BE COMPLETED ONLY ONCE on your first HIT.**
2. **NON-ALPHABETICAL CHARACTERS**: Do not use any non-alphabetical character, e.g. commas, hyphens, exclamation marks etc. But include apostrophe in appropriate places, e.g. " don't ", " can't "
3. **NUMBERS AND SPECIAL SYMBOLS**: Write the English/Spanish word, not the number. For example, write "eighty percent" and not "80%", write "seventy degrees fahrenheit" and not "70 F". In short write everything in the English alphabet.
4. **Please, pay attention in rating the audio characteristics and translation quality (acoustic and lexical). You may need to hear the audio clips multiple times to rate accurately if the emphasis is conveyed in a similar manner in both languages.**
5. **Understanding of emphasis is crucial in answering questions 1-4. You may want to review section 2 multiple times during your work in case you need to refresh the emphasis understanding.**
6. There are similar surveys available (Emotions, Emphasis and Intonation). Please complete the surveys you have completed section 2, otherwise your HITS will be rejected. This survey is related to emphasis questions.
7. You can add your comments at the very end.

Section 1: PER BILINGUAL CLIPS QUESTIONS

Please answer all questions (1-6) for each set of English/Spanish audio clips.

English audio:

Spanish audio:

1. Please rate the quality (acoustic and lexical) of the translation

1 (bad) 2 (poor) 3 (fair) 4 (good) 5 (excellent)

2. Are there any words/phrases emphasized in the English Audio?

Yes No

3. How well is the emphasis of the English audio preserved in the Spanish audio?

1 (bad) 2 (poor) 3 (fair) 4 (good) 5 (excellent)

4. What is your confidence in rating the emphasis preservation?

1 (bad) 2 (poor) 3 (fair) 4 (good) 5 (excellent)

5. Please provide the transcription of the English audio

6. Please provide the transcription of the Spanish audio

Fig. 2. The survey used to validate the hypothesis claimed in the paper.

Finally, we asked for 8 surveys filled for each individual utterance pair. In total, 202 different annotators participated in the surveys. 32.6% and 58.7% of the annotators reported English and Spanish language as the native language respectively. The rest reported other languages. We collected 5977 samples from the movies data and 3895 samples from the S2SData. If we define emphasis transfer as giving an emphasis transfer rating above 3, then 78.4% of the S2SData samples and 84.7% of the movies data have been rated with transfered emphasis. After the results had been collected, we computed the correlation and mutual information [20] between the emphasis transfer rating and the quality of the translation.

6. RESULTS AND DISCUSSION

6.1. Perceptual prosodic emphasis experiments

Fig. 3 shows the normalized counts of translation quality given the rating that emphasis was transfered. By normalized counts, we mean the histogram of the translation quality ratings divided by the number of samples. Hence, each column in Fig. 3 represents the normalized counts for each emphasis rating. We plot the distribution per column to remove any bias coming from unequal priors of each rating value as reported in section 5. In the title of each plot, we provide the correlation and mutual information of emphasis transfer across the various levels of the translation quality variable. The lighter color indicates a high normalized count. Both data sets, show very similar trends. However, the movies data set indicates that some times emphasis transfer is rated as "bad" (i.e. rating 1) but still we get good translation quality rating. That might happen because, in some cases, although conceptually identical, may be paraphrased significantly to make emphasis transfer comparisons difficult (note the temporal synchronicity and potential lip-syncing constraints placed on actors).

	S2SData	Movies
Confidence$\geq$4	0.52	0.49
Confidence$\geq$4, No Emph. present	0.46	0.48
Confidence$\geq$4, Emph. present	0.54	0.50

Table 1. Correlation coefficient when the results are conditioned on the confidence of the annotators and on the cases whether there exists emphasis in the English utterance.

Correlation gives a comparison tool to judge whether there is a linear or inversely linear relation between the emphasis transfer and the quality of the speech translation. The lighter color on the diagonals in Fig. 3 indicates that the hypothesis that more faithful transfer of prosodic emphasis is correlated with perceived overall translation quality; this is validated at the 0.54 correlation level between the emphasis transfer and the quality of the speech translation for the S2SData set and 0.53 for the Movies data set. All results are significant against the no correlation hypothesis using a t-test at p-value=0.01. For any non-linear relations, mutual information is used which is a measure of the predictive power between the two variables of interest and for both data sets the mutual information between the emphasis transfer and the quality of the speech translation is 0.19.

Table 1 presents the correlation given that the confidence of the annotators is greater than 3. Similarly, we present the correlation given that the confidence of the annotators is greater than 3 and they perceived an emphasized word/phrase in the English side or they did not perceive an emphasized word/phrase. Results show that the hypothesis is validated at the 0.46-0.54 correlation level even when annotators report confidence greater than 3 with or without the presence of emphasized words for both S2Sdata and Movies data sets. All results are significant at p-value=0.01.

As reported in section 5, there is a bias towards samples that have been rated as prosodic emphasis transfered. To eliminate any effect of this bias, we randomly picked 500 samples from each class (One class contains the points where emphasis transfer rating is greater than 3 and the rest points are in the other class) for each data set and computed the correlation coefficient. This experiment was repeated 1000 times with replacement. The average correlation of this experiment is 0.54 for S2SData and 0.51 for movies.

It is also interesting to examine the correlation values across annotators. Fig. 4(a) indicates that the main mass of the correlation between the quality of the translation and prosodic emphasis transfer is around 0.5 as expected from the overall correlation figures. In this histogram, the two data sets are reported together. The median point on the histogram is 0.52 which is close to the overall correlation coefficient reported. Fig. 4(b) shows a scatter plot of the number of samples annotated by each annotator and the corresponding correlation coefficient. Annotators with negative correlation have annotated

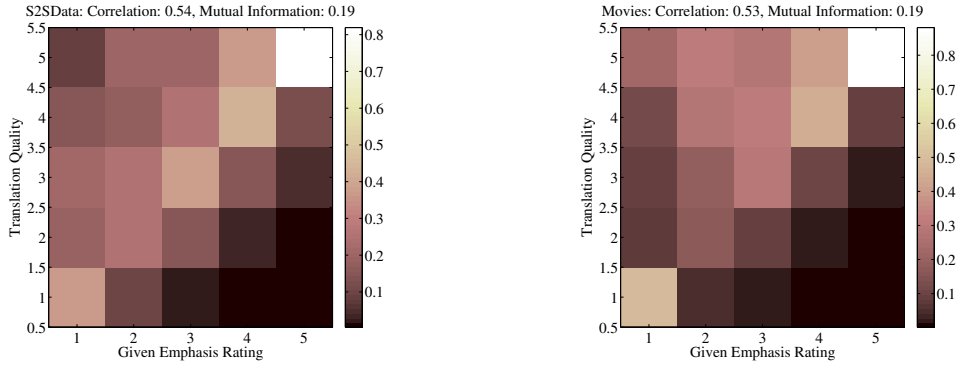


Fig. 3. shows the normalized counts (normalized histogram) of translation quality given the rating that emphasis was transfered. Thus, each column sums up to 1 and represents the distribution of the translation quality for each emphasis transfer rating for both the S2Sdata and Movies data set.

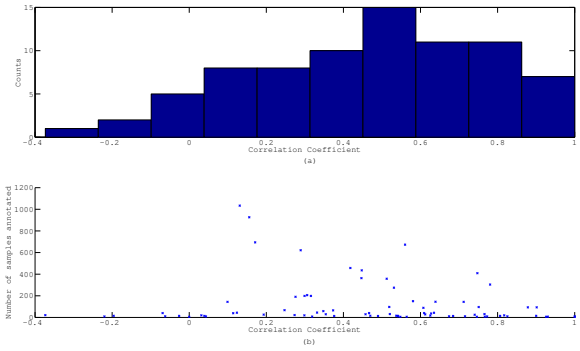


Fig. 4. Fig. 4(a) shows the histogram of correlation between the quality of the translation and prosodic emphasis transfer. Fig. 4(b) shows the scatter plot of the number of samples completed by an annotator vs the correlation between the quality of the translation and prosodic emphasis. In both cases, we included annotators with more than 5 samples.

very few samples and their effect was minimal on the overall correlation score. Also, a few people that annotated a lot of samples (e.g. above 700 samples) gave average correlation of 0.14 which is well below the overall median. However, the average completion time per sample of annotators having above 700 samples is 58.5 seconds much lower than the global average of 98.6 seconds which questions those annotators quality.

Overall, we conclude that there is approximately 0.5 correlation between the emphasis transfer and the quality of the speech translation. However, to make a stronger statement with higher correlation we might need to include other prosodic variables, for example, intonation, emotional state, etc.

6.2. Lessons learned

From our experience with the S2S subjective experiments on Amazon Turk, we learned that it is important to have a training part, mandate annotators to take the training part and have a way to validate that they went through this training procedure. Initially, we did this experiment on a small scale with written instructions but without the training part and many annotators were asking questions about emphasis and what we expect from them. After manually creating clear examples on what we mean by prosodic emphasis the questions on this topic were minimal.

Apart from a well prepared training procedure and explanation, it is important to evaluate the annotators understanding of both languages. Initially, we had only the questions $1-4$ in the survey (Fig. 2) and soon realized that we were getting bad annotations (completed extremely fast) from people that we couldn't say if they are fluent in both languages or not. So we added the questions $5-6$ and mandated the annotators to transcribe all utterances in both languages. This helped us to ensure that annotators were actually listening to the samples they were rating and also we filtered a lot of annotators that were not fluent in both languages.

Finally, Amazon Turk provides no procedure to limit the number of samples annotated per person (only ensures that annotators are not presented with the same sample more than once). This created imbalances in number of the samples annotated per person. To limit such imbalances one has to request annotators to stop after a certain upper bound number of annotations and if they do not comply, then exclude them from the task.

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we have presented a perceptual study to establish the hypothesis that there is a relation between the emphasis transfer and the quality of speech translation. The hypothesis is validated at 0.53-0.54 correlation level on the two data sets used. The results are significant at p-value=0.01. We also discussed the lessons we learned in rating perceptually the S2S translation quality in these subjective experiments using Amazon Turk.

Some future directions we want to investigate include expanding this work in carrying out the experiments in other language pairs. In addition, we want to expand the study to the relation of speech translation quality and other paralinguistic cues transfer, for instance, intonation and emotions.

8. 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] Narayanan S., Ananthakrishnan S., Belvin R., Ettaile E., Ganjavi S., Georgiou PG, Hein CM, Kadambe S., Knight K., Marcu D., et al., "Transonics: A speech to speech system for english-persian interactions," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 670–675.

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

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

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

[9] 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.

[10] Campbell N., "On the use of nonverbal speech sounds in human communication," *Verbal and nonverbal communication behaviours*, pp. 117–128, 2007.

[11] Crystal D., *Prosodic systems and intonation in English*, vol. 1, Cambridge University Press, 1976.

[12] Brazil D. et al., *Discourse Intonation and Language Teaching.*, ERIC, 1980.

[13] Callison-Burch C., "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 286–295.

[14] Mohammad S.M. and Turney P.D., "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 26–34.

[15] Kunath S.A. and Weinberger S.H., "The wisdom of the crowd's ear: speech accent rating and annotation with amazon mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 168–171.

[16] Ambati V., Vogel S., and Carbonell J., "Active learning and crowd-sourcing for machine translation," *Language Resources and Evaluation (LREC)*, vol. 7, pp. 2169–2174, 2010.

[17] R. Belvin, W. May, S. Narayanan, P. Georgiou, and S. Ganjavi, "Creation of a doctor-patient dialogue corpus using standardized patients.," *In Proc. LREC, Lisbon, Portugal*, 2004.

[18] A. Tsiartas, P. Ghosh, P. G. Georgiou, and S. Narayanan, "Bilingual audio-subtitle extraction using automatic segmentation of movie audio," in *the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2010, pp. 5624–5627.

[19] McCowan, A. Iain, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," Idiap-RR Idiap-RR-73-2004, IDIAP, Martigny, Switzerland, 0 2004.

[20] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory (2nd Edition)*, Wiley-Interscience, 2006.