Factored translation models for enriching spoken language translation with prosody

Vivek Kumar Rangarajan Sridhar[1], *Srinivas Bangalore*[2], *Shrikanth Narayanan*[1]

[1]Speech Analysis and Interpretation Laboratory, University of Southern California
[2]AT&T Labs - Research

vrangara@usc.edu, srini@research.att.com, shri@sipi.usc.edu

10.21437/Interspeech.2008-675

Abstract

Key contextual information such as word prominence, emphasis, and contrast is typically ignored in speech-to-speech (S2S) translation due to the compartmentalized nature of the translation process. Conventional S2S systems rely on extracting prosody dependent cues from hypothesized (possibly erroneous) translation output using only words and syntax. In contrast, we propose the use of factored translation models to integrate the assignment and transfer of pitch accents (tonal prominence) during translation. We report experiments on 2 parallel corpora (Farsi-English and Japanese-English). The proposed factored translation models provide a relative improvement of 8.4% and 16.8% in pitch accent labeling accuracy over the post-processing approach for the two corpora respectively.

1. Introduction

Current speech translation approaches predominantly rely on a pipeline model wherein a speech recognizer transcribes the source language speech into text. Typically, the 1-best ASR hypothesis text is considered for machine translation followed by synthesis into speech in the target language. Such an approach loses the rich information contained in the source speech signal that may be vital for successful communication. Disregarding such information may result in ambiguous concept transfer in translation (e.g., providing improper utterance chunking; erroneously emphasizing a target language word or phrase). In other cases, key contextual information such as word prominence, emphasis, and contrast can be lost in the speech-to-text conversion. In this paper, we investigate issues related to accurate capture and transfer of prosodic information – properties that signify aspects of intonation, phrasing, rhythm and emphasis.

Prosodic information has mainly been used in speech translation for utterance segmentation [1, 2] and disambiguation [3]. The VERBMOBIL speech-to-speech translation (S2S) system [3] utilized prosody through clause boundaries, accentuation and sentence mood for improving the linguistic analysis within the speech understanding component. The use of clause boundaries improved the decoding speed and disambiguation during translation. More recently Aguero et al. [4] have proposed a framework for generating target language intonation as a function of source utterance intonation. They used an unsupervised algorithm to find intonation clusters in the source speech similar to target speech. However, such a scheme assumes some notion of prosodic isomorphism either at word or accent group level.

In this work, we incorporate prosodic prominence (represented through categorical *pitch accent* labels) in a statistical speech translation framework by injecting these labels into the target side of translation. Our approach generates enriched tokens on the target side in contrast with conventional systems that predict prosody from the output of the statistical machine translation using just hypothesized text and syntax. The proposed framework integrates the assignment of prominence to word tokens within the translation engine. Hence, the automatic prosody labeler can exploit lexical, syntactic *and* acoustic-prosodic information. Furthermore, the enriched target language output can be used to facilitate prosody enriched text-to-speech synthesis, the quality of which is typically preferred by human listeners [5]. A system level illustration of the proposed framework in comparison with conventional S2S systems is presented in Figure 1.

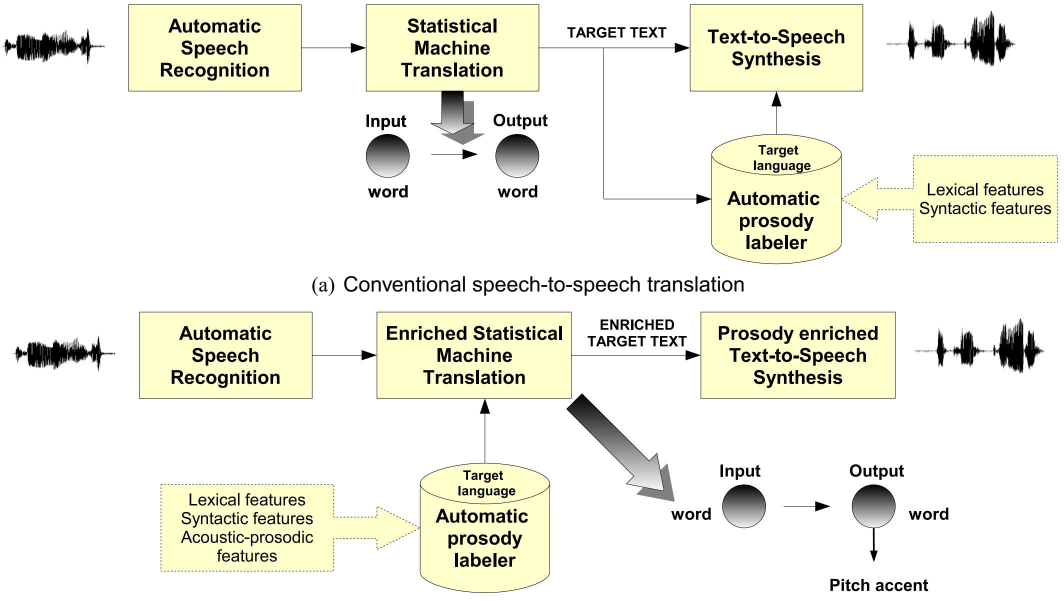
The rest of the paper is organized as follows: Section 2 describes the automatic prosody labeler used in this work. Section 3 contains a summary of the parallel corpora used in the translation experiments. Section 4 formulates the problem and describes the factored translation models used in our experiments. Section 5 summarizes the results of our experiments and Section 6 concludes the paper with a discussion and directions for future work.

2. Automatic prominence labeling

In this section, we describe the classifier used for automatic prominence detection in the rest of the paper. The classifier was trained on a subset of the Switchboard corpus that had been hand-labeled with pitch accent markers [6]. The corpus is based on about 4.7 hours of hand-labeled conversational speech (excluding silence and noise) from 63 conversations of the Switchboard corpus and 1 conversation from the CallHome corpus. The corpus contains about 67k word instances (excluding silences and noise). Prominent syllables were marked only with "*" for indicating a pitch accent (tonally cued prominence) or "*?" for a possible prominence (i.e., uncertainty about presence of a pitch accent). We mapped the pitch accent labels on syllables to words for training a word-level pitch accent classifier with two classes, *accent* and *none*.

We use a maximum entropy model for the prominence labeling. Given a sequence of words w_i in an utterance $W = \{w_1, \cdots, w_n\}$, the corresponding syntactic information sequence $S = \{s_1, \cdots, s_n\}$ (for e.g., parts-of-speech, syntactic parse, etc.), a set of acoustic-prosodic features $A = \{\mathbf{a}_1, \cdots, \mathbf{a}_n\}$, where $\mathbf{a}_i = (a_i^1, \cdots, a_i^{t_{w_i}})$ is the acoustic-prosodic feature vector corresponding to word w_i with a frame length of t_{w_i} and a prosodic label vocabulary ($l_i \, \epsilon \, \mathcal{L}, |\mathcal{L}| = V$), the best prosodic label sequence $L^* = l_1, l_2, \cdots, l_n$ is obtained by approximating the sequence classification problem,

Accepted after peer review of full paper


Automatic Speech Recognition
Statistical Machine Translation
TARGET TEXT
Text-to-Speech Synthesis
Input
Output
word
word
Target language
Automatic prosody labeler
Lexical features
Syntactic features
Automatic Speech Recognition
Enriched Statistical Machine Translation
ENRICHED TARGET TEXT
Prosody enriched Text-to-Speech Synthesis
Target language
Automatic prosody labeler
Lexical features
Syntactic features
Acoustic-prosodic features
Input
Output
word
word
Pitch accent

(a) Conventional speech-to-speech translation

(b) Prosody enriched speech-to-speech translation

Figure 1: Illustration of the proposed scheme in comparison with conventional approaches

using conditional independence assumptions, to a product of local classification problems as shown in Eq.(2). The classifier is then used to assign a prosodic label to each word conditioned on a vector of local contextual features comprising the lexical, syntactic and acoustic information.

$$L^* = \arg\max_L P(L|W,S,A) \quad (1)$$

$$\approx \arg\max_L \prod_{i=1}^{n} p(l_i|w_{i-k}^{i+k}, s_{i-k}^{i+k}, \mathbf{a}_{i-k}^{i+k}) \quad (2)$$

$$= \arg\max_L \prod_{i=1}^{n} p(l_i|\mathbf{\Phi}(W,S,A,i)) \quad (3)$$

where $\mathbf{\Phi}(W,S,A,i) = (w_{i-k}^{i+k}, s_{i-k}^{i+k}, \mathbf{a}_{i-k}^{i+k})$ is a set of features extracted within a bounded local context k. In this paper, the lexical features are word trigrams, syntactic features are trigrams of part-of-speech tags and supertags [7] and acoustic-prosodic features are the normalized (over utterance) f0 and energy values extracted over 10ms frames.

We use the machine learning toolkit LLAMA [8] to estimate the conditional distribution $P(l_i|\mathbf{\Phi})$ using maxent. The 10-fold cross-validation performance of the classifier on the subset of Switchboard corpus described above is presented in Table 1 (chance=67.48%). The pitch accent detection accuracy reported here is close to the state-of-the-art for spontaneous speech from the Switchboard corpus [9]. More details about the automatic prosody labeler can be found in [10].

Cues used (k=3)	Accuracy (%)
	Pitch accent
Lexical	72.68
Lexical+Syntactic	75.90
Prosodic	74.34
Lexical+Syntactic+Prosodic	78.52

Table 1: Pitch accent detection accuracies for various cues on the prosodically labeled Switchboard corpus.

3. Data

We report experiments on two different parallel corpora of spoken dialogs: Farsi-English and Japanese-English. The Farsi-English data used in this paper was collected for doctor-patient mediated interactions in which an English speaking doctor interacts with a Persian speaking patient [11]. The corpus consists of 9315 parallel sentences with corresponding audio for each English sentence. The conversations are spontaneous and the audio was recorded through a microphone (22.5KHz).

The Japanese-English parallel corpus is a part of the "How May I Help You" (HMIHY) [12] corpus of operator-customer conversations related to telephone services. The corpus consists of 12239 parallel sentences with corresponding English side audio. The conversations are spontaneous and the audio was recorded over a telephone channel (8KHz). The statistics of the data corpora are summarized in Table 2.

4. Enriching translation with prosody

In this section, we formulate the problem of using rich prosodic annotations in speech translation. Let S_s, T_s and S_t, T_t be the speech signals and equivalent textual transcription in the source and target language, and L_t the enriched representation (prosody) for the target speech. We formalize our proposed enriched S2S translation in the following manner:

$$S_t^* = \arg\max_{S_t} P(S_t|S_s) \quad (4)$$

$$P(S_t|S_s) = \sum_{T_t,T_s,L_t} P(S_t,T_t,T_s,L_t|S_s) \quad (5)$$

$$\approx \sum_{T_t,T_s,L_t} P(S_t|T_t,L_t).P(T_t,L_t|T_s).P(T_s|S_s) \quad (6)$$

where Eq.(6) is obtained through conditional independence assumptions. Even though the recognition and translation can be performed jointly [13], typical S2S translation frameworks compartmentalize the ASR, MT and TTS, with each component

	Training				Test			
	Farsi	Eng	Jap	Eng	Farsi	Eng	Jap	Eng
Sentences	8066		12239		925		604	
Running words	76321	86756	64096	77959	5442	6073	4619	6028
Vocabulary	6140	3908	4271	2079	1487	1103	926	567
Singletons	2819	1508	2749	1156	903	573	638	316

Table 2: Statistics of the training and test data used in the experiments.

maximized for performance individually.

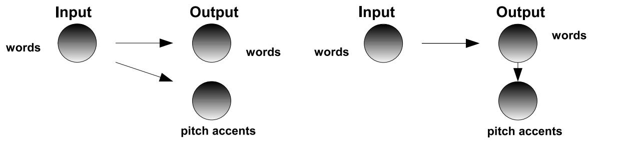
$$\max_{S_t} P(S_t|S_s) \approx \max_{S_t} P(S_t|T_t^*, L_t^*) \times \max_{T_t, L_t} P(T_t, L_t|T_s^*) \times \max_{T_s} P(T_s|S_s) \quad (7)$$

where T_s^* is the output of speech recognition, T_t^* and L_t^* are the target text and enriched prosodic representation obtained from translation. While conventional approaches address the detection of L_t^* separately through postprocessing, we integrate this within the translation process thereby enabling the use of acoustic-prosodic information in training the translation engine (see Figure 1). In this work, we do not address the speech synthesis part and assume that we have access to the reference transcripts.The rich annotations (L_t) can be syntactic or semantic concepts [14, 15] or as in this work, pitch accent labels predicted from the model described in Section 2.

4.1. Factored translation models for incorporating prominence

Factored translation models [15] have been proposed recently to integrate linguistic information such as part-of-speech, morphology and shallow syntax in conventional phrase-based statistical translation. The framework allows for integrating multiple levels of information into the translation process instead of incorporating linguistic markers in either preprocessing or postprocessing. For example, in morphologically rich languages it may be preferable to translate lemma, part-of-speech and morphological information separately and combine the information on the target side to generate the output surface words.

Factored translation models have been used primarily to improve the word level translation accuracy by incorporating the factors in phrase-based translation. In contrast, we are interested in integrating factors such as pitch accent labels in speech translation with the objective of maximizing the accuracy of the output factors themselves. By facilitating factored translation with pitch accent labels predicted from prosodic, syntactic and lexical cues, our enriched translation scheme can produce output with improved pitch accent assignment accuracy. On the other hand, predicting prominence at output of conventional S2S systems is subject to greater error due to typically noisy translations and lack of direct acoustic-prosodic information. Figure 2 illustrates the type of factored models used in this paper.



(a) Factored model 1 (b) Factored model 2

Figure 2: Illustration of the proposed factored translation models to incorporate prominence

Factored model 1 represents joint translation of words and prominence. Thus, the phrase translation table obtained for such a model would have compound tokens (word+prominence) in the target language. However, with a factored approach we can build the alignments based on the words alone, thus avoiding data sparsity typically introduced by compound tokens. Factored model 2 translates input words to output words and generates prominence labels from the output word forms through a generation step.

5. Experiments and Results

The translation experiments reported in this work were conducted using the Moses[1] toolkit for statistical phrase-based translation. We report results for three scenarios which vary in terms of how the prominence labels are produced in the target language.

1. Post processing: The pitch accent labels are produced at the output of the translation block using lexical and syntactic cues from hypothesized text
2. Factored model 1: Factored model that translates source words to target words and pitch accents
3. Factored model 2: Factored model that translates source words to target words which in turn generate pitch accents

Table 3 summarizes the results obtained in terms of BLEU score [16], lexical selection accuracy and prosodic accuracy. Lexical selection accuracy is measured in terms of the F-measure derived from recall ($\frac{|Res \cap Ref|}{|Ref|} * 100$) and precision ($\frac{|Res \cap Ref|}{|Res|} * 100$), where Ref is the set of words in the reference translation and Res is the set of words in the translation output. Prosodic accuracy is defined as $\frac{\#\ correct\ pitch\ accents \in (Res \cap Ref)}{|Res \cap Ref|} * 100$. Figure 3 illustrates the computation of prosodic accuracy for an example utterance.

Source : من من برای آسم دارو مصرف میکنم
Reference : I_none I'm_* taking_none medication_* for_none asthma_*
Hypothesis: I'm_none on_none medication_* for_none asthma_*

Ref ∩ Res : {I'm, medication, for, asthma}
#correct pitch accents: 3

Figure 3: Illustration of the process used to calculate prosodic accuracy

The reference pitch accent labels for the English sentences were obtained from the automatic prominence labeler described in Section 2 using lexical, syntactic and prosodic cues. The language models were trigram models created only from the training portion of each corpus. The results in Table 3 indicate that the assignment of correct pitch accents to the target words improves with the use of factored translation models. Factored

[1] http://www.statmt.org/moses

Translation model	Farsi-English			Japanese-English		
	Lexical F-score	BLEU	Prosodic accuracy	Lexical F-score	BLEU	Prosodic accuracy
Postprocessing	56.46	22.90	74.51	78.98	54.01	68.57
Factored model 1	56.18	22.93	80.83	79.00	54.04	80.12
Factored model 2	56.07	22.85	80.57	78.56	53.97	79.56

Table 3: Evaluation metrics for the two corpora used in experiments (all scores are in %)

model 1 that translates input word forms to output word forms and pitch accents achieves the best performance. We obtain a relative improvement of 8.4% and 16.8% in prosodic accuracy for the two corpora in comparison with the postprocessing approach. In the postprocessing approach, the pitch accent classifier was trained on lexical, syntactic and acoustic-prosodic features from clean sentences, but evaluated on possibly erroneous machine translation output. Furthermore, the lack of acoustic-prosodic information at the output of machine translation results in lower prosodic assignment accuracy. On the other hand, factored models integrate the pitch accent labels derived from lexical, syntactic and acoustic-prosodic features within the translation framework. Thus, the prosodic accuracy obtained is consistently higher than the postprocessing scheme.

Table 3 also illustrates translation performance at the word level. For both the factored translation models, the word-level BLEU score and lexical selection accuracy are close to the baseline model that uses no pitch accent labels within the translation framework. Thus, the improvement in prosodic assignment accuracy is obtained at no significant degradation of the word-level translation performance.

6. Discussion and Future Work

It is important to note that the pitch accent labels used in our translation system are predictions from the maxent based prosody labeler described in Section 2. We do not have access to the true reference labels; thus, some amount of error is to be expected in the predictions. Improving the current prosody labeler and developing suitable adaptation techniques are part of future work.

The models proposed in this work may be especially useful for tonal languages such as Chinese where it is important to associate accurate tones to syllables. Our framework can produce enriched target output by integrating the acoustic-prosodic information during translation in comparison with conventional S2S translation systems that postprocess the output to predict prominence.

While we have demonstrated that our framework can improve the accuracy of prominence labels in the target language, it can potentially be used to integrate any word-level rich annotation dependent on acoustic-prosodic features (e.g., boundary tones, emotion, etc.). We have not used optimization techniques such as minimum error rate training (MERT) in this work due to the relatively small size of the corpora. The use of such techniques could potentially lead to further improvements. Finally, the proposed framework needs to be evaluated by including a speech synthesis system that can make use of prosodic markers. We plan to address this also as part of future work.

7. Acknowledgments

This research has been supported by a grant from DARPA.

8. References

[1] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation with automatic boundary prediction," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[2] C. Fügen and M. Kolss, "The influence of utterance chunking on machine translation performance," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[3] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system," *IEEE Transactions on Speech and Audio processing*, vol. 8, no. 5, pp. 519–532, September 2000.

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

[5] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[6] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 119–121.

[7] S. Bangalore and A. K. Joshi, "Supertagging: An approach to almost parsing," *Computational Linguistics*, vol. 25, no. 2, June 1999.

[8] P. Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.

[9] M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart, and A. Krasnyanskaya, "Parsing speech and structural event detection," JHU Summer Workshop, Tech. Rep., 2005.

[10] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech and Language Processing*, vol. In press, 2008.

[11] S. Narayanan et. al, "Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains," in *Proc. of ICASSP*, Toulose, France, May 2006.

[12] A. Gorin, G. Riccardi, and J. Wright, "How May I Help You?" *Speech Communication*, vol. 23, pp. 113–127, 1997.

[13] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. of Eurospeech*, 2005.

[14] L. Gu, Y. Gao, F. H. Liu, and M. Picheny, "Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 377–392, March 2006.

[15] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of EMNLP*, 2007.

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," IBM T.J. Watson Research Center, Tech. Rep., 2002.