INTENT TRANSFER IN SPEECH-TO-SPEECH MACHINE TRANSLATION

Gopala Krishna Anumanchipalli[†‡], *Luís C. Oliveira*[‡], *Alan W Black*[†]

[†]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[‡]Spoken Language Systems Laboratory, INESC-ID/IST Lisboa, Portugal
{gopalakr,awb}@cs.cmu.edu, lco@inesc-id.pt

ABSTRACT

This paper presents an approach for transfer of speaker intent in speech-to-speech machine translation (S2SMT). Specifically, we describe techniques to retain the prominence patterns of the source language utterance through the translation pipeline and impose this information during speech synthesis in the target language. We first present an analysis of word focus across languages to motivate the problem of transfer. We then propose an approach for training an appropriate transfer function for intonation on a parallel speech corpus in the two languages within which the translation is carried out. We present our analysis and experiments on English↔Portuguese and English↔German language pairs and evaluate the proposed transformation techniques through objective measures.

Index Terms**—** Speech Translation, Prominence, Focus, Speech Synthesis, Cross-lingual Transfer

1. INTRODUCTION

Intent in speech is manifested in how a sentence is delivered, its phrasing, rhythm, intonation, energy and voice quality. Given the same sentence, speakers have large variability and freedom in focussing any concept (word) they choose to, and the degree to which the emphasis is laid. This prominence pattern of the words in an utterance bears information about the words' relevance, given/newness etc., in addition to the general style of the speaker. These aspects broadly fall under the 'augmentative' and 'affective' parts of prosody, the extra information in speech, to ensure that the intended message is unambiguously decoded by the listeners [1]. In this work, we use the term 'intent' to exclusively refer to such aspects within intonation. While it is ideal for Text-to-speech (TTS) systems to synthesize as appropriate to the underlying meaning of the sentence, intent is largely under-represented in text. However, there are certain domains where this information may be accessible to the synthesizer. In this work, we deal with one such domain, speech-to-speech machine translation (S2SMT). The goal of S2SMT is to take as input, speech in one language and automatically 'dub' it to generate as output, a translated sentence with the same meaning spoken in another language.

Traditional approaches to S2SMT use a pipeline architecture where speech in a source language is passed through an automatic speech recognizer (ASR), the ASR hypothesis is translated to a target language using a machine translation system (SMT). The translation output is then passed on to a TTS system in the target language. Since these individual component systems are still fragile in practice, S2SMT systems have not yet become commonplace. This is partly due to the errors that each system contributes, but also due to the cumulative loss of information along the traditional pipeline. In general, the prosody of the source utterance is discarded by the ASR system, and is not accessible to the TTS system in the target language. This information is critical if S2SMT systems are ever to match the performance of professional translators or dubbing artists.

While there has been considerable work in the ASR, SMT components and in tightening the interface between them to improve speech translation [2] [3] [4], issues for speech synthesis within this framework remain to be studied [5][6]. Previously, prosody in the source side has been used to improve the performance of the ASR systems for verifying different linguistic hypotheses [7]. To integrate the TTS component for S2SMT, Parlikar el al., [6] propose to improve the fluency of the SMT output optimized for TTS. There are also techniques for cross-lingual conversion of spectral information that the TTS can be employ to match the original speaker's voice after translation [8] [9] [10].

In this work, we want to further exploit the source prosodic information by imposing it appropriately on the target side after translation, consequently transferring the speaker intent across in S2SMT and truly 'complete' the goal of translation. Figure 1 situates the current problem within the framework of speech translation.

While beyond the scope of this paper, Fig. 1 also shows other outstanding problems at the source-target interface, i.e., speaker identity and sentence boundaries (relevant for better audio-visual synchronization in automatically dubbed videos). In this paper, we only deal with transfer of speaker intent as conveyed through intonation. We address this problem by learning how the intonational correlates of focus change across languages, considering the case of two language pairs for translation in this work. In Section 2, we list

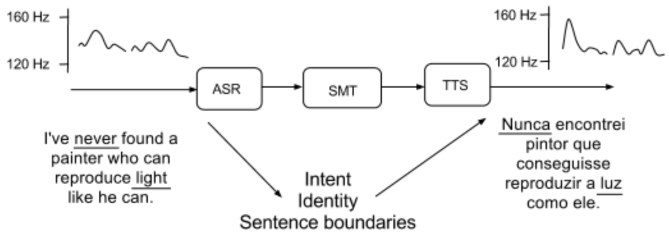
160 Hz
120 Hz
ASR
SMT
TTS
160 Hz
120 Hz
I've never found a painter who can reproduce light like he can.
Intent
Identity
Sentence boundaries
Nunca encontrei pintor que conseguisse reproduzir a luz como ele.

Fig. 1. Schematic illustration of the proposed intent transfer technique within the S2SMT framework, here for English→Portuguese.

the resources we use in this work, including a new English-Portuguese parallel speech corpus we created for this language pair. Section 3 presents some manual and automatic analysis of focus for motivating the current problem. Approaches for training of cross-lingual accent transformation functions is presented in Section 4, followed by objective evaluation of the proposed methods in Section 6. Our results show that the approach effectively transfers the word prominence patterns cross-lingually.

We reiterate that correlates of focus also exist in the energy, duration and phrasing patterns around the associated concepts. In this work however, we deal only with the intonational aspects, manifested as appropriate pitch accents to convey word focus.

2. TOOLS AND RESOURCES

In order to computationally model intent transformation across languages, it is essential to first analyze any systematic patterns in how intonation is employed to convey intent in the two languages considered. This can be tractably done using a parallel speech corpus, where sentences with the same underlying meaning (parallel text) in the two languages are recorded by the same speaker fluent in both the languages, preferably by a bilingual speaker. The premise here is that since the underlying meaning and speaker are the same in both languages, speaker intent and intonation are also comparable. Since such resources do not readily exist for the English-Portuguese (`en-pt`) language pair, here we describe the construction of such a corpus.

2.1. Parallel Speech Corpora

As the text corpus from which to record, we use *UP*, the in-flight magazine of Portugal's *TAP* airlines. The magazine has parallel articles, parallel at the paragraph and sentence levels, on a variety of topics including travel, cuisine, art and culture. This ensures a good coverage of proper nouns and syntactic constructions in the recording prompts, suited for training high-quality natural-sounding TTS systems. From a vast collection of articles, an optimal set of paragraphs (optimized for phonetic coverage) is chosen to be recorded by a native Portuguese speaker fluent in both the languages. The choice of recording at the paragraph level was deliberately made to give the speaker enough linguistic context for employing natural prosody, which is otherwise difficult to elicit in sentence level read speech recordings. The recordings were done alternatively for each paragraph, first in Portuguese and then in English, so that the speaker is likely to employ the same intent in the two languages. However the speaker is not given explicit instructions to maintain the same focus/prominence patterns in the two instances of recordings.

These paragraph level utterances are automatically chunked at the sentence level and are phonetically segmented using the `islice` module [11] within Festvox voice building suite. The duration of the speech is approximately 1 hour in each language. The corpus statistics are as presented in the Table 1.

Table 1. Statistics of the English-Portuguese parallel speech corpora

Language	English	Portuguese
#Paragraphs	84	84
#Sentences	420	420
#Tokens	8184	8211
#Words	2934	3283
#Tokens/Sentence	19.49	19.55
Duration(in mins)	60.36	59.47

Additionally, we also present results of automatic focus analysis on an English-German (`en-de`) parallel speech database generously provided by the EMIME project [10]. The statistics of this corpus is presented in Table 2.

Table 2. The EMIME English-German parallel speech corpus

Language	English	German
Speaker ID	GM1	GM1
#Paragraphs	—	—
#Sentences	145	145
#Tokens	1301	1198
#Words	763	697
#Tokens/Sentence	8.97	8.26
Duration(in mins)	11.68	11.87

The EMIME databases are read speech recordings at the sentence level. Note that the #Tokens are higher in number and more comparable in the `en-pt` language pair since it is more free style magazine content, and due to the fact that German is agglutinative. Also note that the `en-de` corpus is much smaller in data size per speaker.

2.2. Word Alignment through Statistical Machine Translation

In comparing intonation of two speakers within a language, prosody is studied across the same linguistic entities (words/ phrases etc). On similar lines, it is necessary to determine comparable linguistic anchors for comparing prosody across languages. To study the correspondence in intonation, we obtain the mapping between the words in the source and target language sentences. We use GIZA++ [12] tool to align the sentences within each language pair. A word alignment model trained on the parallel text in these databases, seeded with the respective Europarl data [13] in these languages is used to obtain the word mappings between the languages. This word alignment information is necessary both in the analysis and synthesis phases.

2.3. Detection of Word Accentedness

Since we are interested in the analysis of intonation, it is necessary to identify the words in the speech utterance that are likely to have a pitch accent on them (a salient excursion on the F0 contour). We use the accentedness detection module of AuToBI [14], which gives a probability of the presence of an intonational accent given the speech waveform and the word boundaries of interest. Although AuToBI also outputs a predicted ToBI category label for each accent it detects, we do not make use of this information. Additionally, we use the the same accentedness models (trained on conversational American English) for all the languages involved. While this may be suboptimal, most acoustic realizations of salient pitch accents are universal enough to be detected, the phonology (and hence the category label) however may be language specific. Since we are interested in focus conveyed through intonation, we use the accentedness detector and not an explicit prominence detector. This is because correlates of focus/prominence also include other prosodic components like duration, which we do not deal with in this work.

2.4. Statistical Parametric Text-to-Speech Synthesis

TTS is the chief component responsible to synthesize the output waveforms with appropriate intonation contours. This demands a flexible speech synthesizer that can take specifications (eg., of word prominence) into account for synthesis. Statistical Parametric Speech Synthesis (SPSS) [15] is the paradigm that renders this flexibility to the TTS systems. We train Clustergen [16] voices, an instance of an SPSS framework for all the languages reported here. Explicit statistical intonation models based on SPAM [17] are trained to enable synthesis of natural sounding and affective intonation contours.

Conventional intonation models in SPSS are built at the level of frame (5-10 ms), the phoneme state level or at the syllable level. In this work, however, we employ a word-level intonation model to make the comparisons in this work more tangible. This makes the setup consistent with the word alignment information output by the SMT system. Given training data, the intonation contours are parameterized as TILT [18] vectors over each word, to quantitatively describe the intonation as a tuple of 4 values (F0 peak, duration, tilt amplitude and tilt). The statistical model involves clustering these vectors with relevant linguistic information, saved as CART trees optimized or a held-out set.

3. CROSS-LINGUAL ANALYSIS OF INTENT

To empirically investigate the relevance of the current problem of cross-lingual intent transfer, in this section we report analysis on a subset of data from the `en-pt` parallel speech corpus.

3.1. Manual Analysis of Cross-lingual Focus

A random set of 75 sentences (about 10 minutes of speech) is chosen and annotated for focus by a trained linguist, fluent in both the languages. The annotator was asked to mark all the focussed word(s) in each sentence. The annotations for each language were carried out in different sessions so as to limit the influence on perception of language stimuli on one another. These annotations are of explicit focus, hence could also include perceived emphasis through energy and duration cues. Table 3 summarizes the focus annotations in both the languages.

Table 3. Results of manual annotation of focus in parallel speech

Language	Total #words	focussed words	non-focussed words	#focussed/ sentence
English	1569	298	1271	3.97
Portuguese	1585	285	1300	3.8

It is no surprise that the expert annotator marked comparable number of words as focussed in either language. To further analyze how much agreement there is, in focussed words across languages, we use SMT alignments between the parallel sentences. Of the 75 sentence subset, alignments were generated only for 1110 `en-pt` word pairs. These also include many-to-one and one-to-many mappings between the words. Among the 1110 word pairs, 336 were marked with focus in the English word and 303 were marked as focussed on the Portuguese word. The intersection between the marked focussed words (focus on both words in pair) is found to be in 145 word pairs (about 48% match). This result is worse than, yet comparable to inter-annotator agreement of prominence on the same set of speech stimuli within a language [19].

It is therefore clear that there is a substantial amount of overlap in the relative prominence across the two languages in the `en-pt` task. This number is still an underestimate, given the mis-alignments from SMT output and the tough nature of the task (paragraph level recordings, without explicit instruction to maintain similar prosody). This analysis reinforces our conjecture that word focus is similar across languages for the same sentence. Hence this information, if available, must be exploited to improve S2SMT.

3.2. Automatic Cross-Lingual Accent Analysis

In order to efficiently and scalably establish the above result without manual annotation, we employ the accentedness detection module of AuToBI, on each language pair in both directions. The output of this module is an array of probabilities of having an accent, one for each word of the sentence. Though these values are just binary class probabilities of weather the word has an accent or not (as opposed to the relative degrees of emphasis among the words), the value does capture prominence patterns, by identifying non-focussed words as such. Moreover, this measure is an attribute only of the intonation contour and not the other prosodic correlates of prominence, in line with the goals in the current work. The accent probability array is sorted and only values ≥ 0.5 are considered as genuinely accented.

The same subset of manually analyzed data is used for the analysis presented in the `en-pt` case. We also report the numbers for `en-de` language pair. Table 4 shows the coverage of the word with the highest accent probability on the source side among the n-best accented words on the target side.

Table 4. % match in automatically detected accented words

task	% accented source words seen in target			
	1-best	2-best	3-best	4-best
en-pt	17.09	32.05	47.43	59.40
pt-en	26.14	41.17	56.86	66.01
en-de	13.51	31.08	47.97	61.48
de-en	4.81	17.64	25.13	32.62

This shows that accented words in the source language are also likely to have an accent on the target side. Recall the high #Tokens/sentence ratios for the `en-pt` databases (Table 1), thus making the 4-best measure quite respectable. The relatively lower numbers in case of `de-en` are possibly due to the issue of compounding in German, where each focussed word maps to more than one, possibly non-focussed words.

We have thus seen compelling evidence to the consistency of prominence patterns across languages.

4. CROSS-LINGUAL INTONATION TRANSFORMATION

In this section we present an approach for conversion of intonation from one language to another. Given a natural utterance in the source language, the goal here is to predict an appropriate intonation contour for the TTS system to synthesis the translated sentence in the target language.

We are motivated by conventional approaches to voice conversion which use parallel data of spectral vectors to train a transformation function between two speakers. Toda et al., [20] employ an approach using Gaussian mixture models of the joint spectral vectors of source and target speakers. Maximum likelihood estimation is used for estimating the spectra of novel sentences of the target speaker, given only the source speech. We use the same framework in our setting — word level TILT vectors are used as the features, and word alignment information from SMT is used to create the parallel data. However, since word focus is primarily on content words, the parallel data is constructed only from the accents over content words. Also, a threshold is determined on accent probability to remove non-accented words from the parallel data. This data is used for training the conversion function between the accent vectors of the source and target language.

Let x_t and y_t be the TILT accent vectors for the corresponding words in both the languages. The joint probability density of the source and target vectors is modelled as the following GMM -

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)})$$

where z_t is the joint accent vector $\begin{bmatrix} x_t' \\ y_t' \end{bmatrix}$, with the GMM having M mixtures with a mean, covariance and mixture weight of the m'th Gaussian component denoted by w_m, $\mu_m^{(z)}$ and $\Sigma_m^{(z)}$ respectively. The Covariance matrix $\Sigma_m^{(z)}$ is constrained to be of the form $\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$, where each partial covariance matrix is set to be a full matrix, because some TILT parameters (eg., duration and tilt amplitude) are positively correlated [21].

The trained function can be used on an novel source utterance's accents along with the translation and the word alignment information to predict an intonation contour with appropriate prominence patterns as used in the original speech. At synthesis time in the target language, the default word-level intonation models predict a TILT vector for each word of the translated sentence. For the content words translated, the associated TILT vector of the original utterance $x(t)$ are converted to $\hat{y_t}$, using the trained conversion function, overriding the default predicted word TILT vectors. This is given by —

$$\begin{aligned} \hat{y_t} &= \sum_{i=1}^{M} p(m_i|x(t), \lambda^{(z)}) E(y_t|x_t, m_i, \lambda^{(z)}), \\ E(y_t|x_t, m_i, \lambda^{(z)}) &= \mu_i^{(y)} + \Sigma_i^{(yx)} {\Sigma_i^{(xx)}}^{-1} (x_t - \mu_i^{(x)}), \\ p(m_i|x(t), \lambda^{(z)}) &= \frac{w_i \mathcal{N}(x_t; \mu^i, \Sigma_i^{(xx)})}{\sum_{j=1}^{M} w_j \mathcal{N}(x_t; \mu_j^{(x)}, \Sigma_j(xx))} \end{aligned}$$

5. EXPERIMENTAL SETUP

Clustergen synthetic voices are built for all the databases within the `en-pt` and `en-de` parallel speech data. Each voice has CART tree models for spectral and duration information. The word-level TILT intonation models are used as the intonation models. These voices are used as the baselines to compare the proposed method against. Essentially, the baselines are standard state-of-the-art TTS systems that only use the text input of the translated sentences.

As the test data, we set aside 10% of the sentences in the target language. We try to objectively measure the distance between the predicted intonation contours for the translated sentences from the reference intonation contours of the test set. We use the conventionally used root mean squared error (`rmse`) and correlation (`corr`). To enable this, the same durations as employed in the reference sentence are employed in synthesis of the test set.

As the proposed intonation model, we use a fusion of the predicted word level intonation model and the transformation model using the joint density GMM on the source utterance accent vectors. For all the function words in the translated sentence, the default predicted word level contour is retained. For the content words, the default is contour linearly interpolated with the transformed intonation contour with a simple mixing weight as given by,

$$F0_{fused} = (\phi)F0_{\mathrm{word}tilt} + (1-\phi)F0_{\mathrm{GMM}vc}$$

where $F0_{\mathrm{word}tilt}$ is the default word level predicted intonation contour, $F0_{\mathrm{GMM}vc}$ is the contour after applying the conversion function on the source utterance's accent vectors and $0 \leq \phi \leq$ is the interpolation weight. This was empirically determined to be 0.6 on a development set across language pairs. The fusion is done to improve the coherence of intonational accents, that could otherwise get effected if the conversion method is directly used, since the technique context insensitive. Table 5 compares the proposed and the baseline intonation contours using the `rmse` and `corr` measures.

Table 5. Objective comparison of synthesized F0 contours

Lang Pair	Default		Proposed	
	`rmse`	`corr`	`rmse`	`corr`
en-pt	17.60	0.51	16.59	0.54
pt-en	15.90	0.47	15.30	0.49
en-de	11.93	0.54	10.98	0.51
de-en	10.27	0.46	10.17	0.46

It can be seen that the proposed method generates intonation contours much closer (lesser `rmse` and higher `corr`)to the reference than the baseline prediction that doesn't exploit the source language prosody. It is also consistently effective in all language pairs, although the degree of improvement is understandably different. To further illustrate the performance of method proposed, Figure 2 shows the predicted intonation contours for three differently emphasized input Portuguese utterances of the sentence *'A lanterna é uma boa invenção'*. The three utterances are varied in which word, the emphasis is laid from among the three content words. In this illustration, the same durations of the baseline system are used across the three utterances for better visualization.

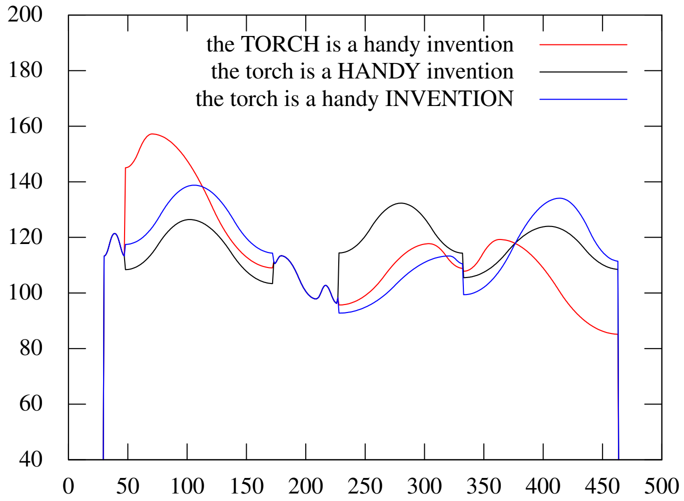

Fig. 2. Synthesized F0's for differently focussed inputs of the Portuguese sentence *'A lanterna é uma boa invenção'*

It can be seen that the synthesized intonation contours in English are also varied to reflect the same prominence patterns as the input. This is quite elegant compared to default TTS systems that invariably produce the same intonation contours for all intents of the underlying text.

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we have motivated the problem of cross-lingual conversion of intent, with respect to intonation in speech. We also created a parallel speech database for the English-Portuguese language pair that is publicly released with this work. We have presented analysis of word focus on two language pairs and proposed an automatic transformation technique of intonational accents. We have objectively shown the improvement of TTS intonation contours employing the proposed techniques.

We are parallelly developing techniques for duration and phrasing that can also exploit the source utterance prosody. These techniques are being tested on translation of Ted talks, based on a collection of lecture style speech data. One challenge for the future is evaluation of different S2SMT systems, which is yet to be addressed. Other interesting directions we wish to pursue is the application of these techniques on other language pairs and multiple bilingual speakers (for extending this work towards a speaker-independent translation) provided in the EMIME dataset. It will be interesting also to characterize the current problem with respect to the various combinations of linguistic and prosodic typologies within the translation language pairs. Another formidable challenge is to make these techniques degrade gracefully in the presence of errors from ASR or SMT systems, both non-trivial as the input speech becomes more spontaneous and free-style.

7. ACKNOWLEDGEMENTS

The authors thank Hugo Meneido for recording the `en-pt` parallel speech corpus and the sentences used in the illustrations; Tiago Luís for generating text alignments and Helena Moniz for focus annotations. Support for this research was provided by the Fundação para a Ciência e a Tecnologia through the CarnegieMellon|Portugal program under grant SFRH/BD/33510/2008, and through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007.

8. REFERENCES

[1] Paul Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[2] Y. Al-Onaizan and L. Mangu, “Arabic ASR and MT integration for GALE,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, april 2007, vol. 4, pp. IV–1285 –IV–1288.

[3] Nicola Bertoldi, Richard Zens, Marcello Federico, and Wade Shen, “Efficient speech translation through confusion network decoding,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 8, pp. 1696–1705, 2008.

[4] M. Wolfel, M. Kolss, F. Kraft, J. Niehues, M. Paulik, and A. Waibel, “Simultaneous machine translation of German lectures into English: Investigating research challenges for the future,” in *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, 2008, pp. 233–236.

[5] P.D. Aguero, J. Adell, and A. Bonafonte, “Prosody generation for speech-to-speech translation,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, may 2006, vol. 1, p. I.

[6] Alok Parlikar, Alan W Black, and Stephan Vogel, “Improving speech synthesis of machine translation output,” in *Interspeech*, Makuhari, Japan, September 2010, pp. 194–197.

[7] E. Noth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, “Verbmobil: The use of prosody in the linguistic components of a speech understanding system,” *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 5, pp. 519–532, 2000.

[8] Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda, “State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis,” in *Interspeech 2009*, Brighton, U.K.,, 2009.

[9] Gopala K. Anumanchipalli and Alan W Black, “Adaptation techniques for speech synthesis in under-resoured languages,” in *Spoken Language Technologies for Under-resoured languages*, Penang, Malaysia, 2010.

[10] Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimki, Reima Karhila, Simon King, Hui Liang, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi jian Wu, and Junichi Yamagishi, “Personalising speech-to-speech translation in the EMIME project,” in *ACL 2010*, 2010.

[11] Kishore Prahallad and Alan W. Black, “Segmentation of Monologues in Audio Books for Building Synthetic Voices,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 5, pp. 1444–1449, 2011.

[12] F. J. Och and H. Ney, “Improved statistical alignment models,” in *ACL 2000*, Hongkong, China, October 2000, pp. 440–447.

[13] Philipp Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proceedings of Machine Translation Summit*, Phuket, Thailand, September 2005, pp. 79–86.

[14] Andrew Rosenberg, “AuToBI - A Tool for Automatic ToBI Annotation,” in *Interspeech 2010*, Chiba, Japan, 2010.

[15] Heiga Zen, Keiichi Tokuda, and Alan W Black, “Review: Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, pp. 1039–1064, November 2009.

[16] Alan W Black, “Clustergen: A statistical parametric synthesizer using trajectory modeling,” in *Interspeech 2006*, Pittsburgh, PA, 2006.

[17] Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W Black, “A Statistical Phrase/Accent Model for Intonation Modeling,” in *Interspeech 2011*, Florence, Italy, 2011.

[18] P. Taylor, “Analysis and synthesis of intonation using the tilt model,” *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.

[19] Yoonsook Mo, Jennifer Cole, and Eun-Kyung Lee, “Naïve listeners’ prominence and boundary perception,” in *Speech Prosody*, 2008.

[20] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222 –2235, nov. 2007.

[21] P Taylor, “Analysis and synthesis of intonation using the tilt model,” *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.