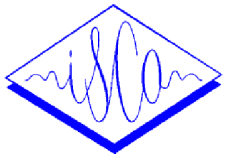

Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

Prosody generation in the Speech-to-Speech Translation Framework

Pablo Daniel Agüero, Jordi Adell and Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain

10.21437/SpeechProsody.2006-118

Abstract

This paper deals with speech synthesis in the framework of speech-to-speech translation. Our current focus is to translate speeches or conversations between humans so that a third person can listen to them in its own language. In this framework the style is not written but spoken and the original speech includes a lot of non-linguistic information (as speaker emotion). In this work we propose the use of prosodic features in the original speech to produce prosody in the target language. Relevant features are found using an unsupervised clustering algorithm that finds, in a bilingual speech corpus, intonation clusters in the source speech which are relevant in the target speech. Preliminary results already show a significant improvement in the synthetic quality.

1. Introduction

Prosody is one of the most important components of human spoken communication. A correct prosody in a text-to-speech system contributes to a better quality in terms of intelligibility, naturalness and pleasantness. It carries different kinds of information. Some of it is strongly related with the text itself (e.g. syntactic structure, semantic disambiguation) but a significant part is not (e.g. emotional, intention). Then, it is not possible to generate such prosody capable to transmit it only from text.

Most widely used speech synthesis applications assume a reading style. Spoken communication is different from written communication. When a human writes a text he is aware that all information will be contained in words themselves. However, when speaking, humans use additional acoustic cues in order to transmit more information. Prosody is thus even more relevant in spoken communication than in written text. Therefore, a greater effort must be done in prosody modelling when dealing with spoken rather than reading style speech synthesis.

The present work has been done within the framework of TC-STAR[1], an EU-funded project focused on speech-to-speech translation (S2ST). In such framework, speech is recognised in a language (source language), then the text is translated into another language (target language) and finally synthesised. The main task of TC-STAR is translation of the European Parliament. In such application, as in translation of broadcast news or lectures, the speaker is not addressing a machine but to people. So, he will not adapt his speaking style to the machine. His speech is produced to be *listened* rather than to be *read*. Therefore, speech will contain lot of information in prosody and thus, it must be *translated* in order to transmit the whole meaning of speech.

There are two possible approaches to this problem. The first one is to make explicit models of each use of prosody, as different emotions, emphasis, speaker attitude, etc. These models have to be used to detect the prosody in the source speech and to generate the correct prosody in the target language. This approach requests for modelling emotions, also for semantic and common knowledge representation, etc. Even if these models could be widely used nowadays it would still be hard to combine them in order to generate a single prosodic model. The second approach is to model the prosodic *translation* implicitly, without any understanding of what it transmitted. In this paper we present an approach to translate the prosodic features based on finding correspondences between features across languages. The main acoustic parameters related with prosody are: *intonation, pauses*, *duration*, *energy* and *voice quality.* However, in the present work this method is only applied to model intonation as a first approach to prosody translation.

This paper is organised as follows. In Section 2 we will describe the prosody translation technique proposed here, together with the intonation model chosen for the experiments . Then, in Section 3 the training algorithm is described in detail. Afterwards, some experiments are presented and results analysed in Section 4. Finally, conclusions are discussed in Section 5.

2. Prosody Translation Model

The basic idea in this work is to label the intonation in the *source speech*, i.e., the speech produced by the speaker in the source language, and use the labels as additional feature to generate the intonation of the *target speech*, i.e., the synthetic speech produced in the target language. The labelling will be based on acoustic measures on the pitch contour, without giving any interpretation to that movement.

The underlying hypothesis is that some paralinguistic information is correlated with the F0 contour. This has already been established in many studies related to emphasis, speaker emotions and even the audience and speaking style (see for instance [1, 2]). In some cases (e.g. some emotions) the F0 movements are in some extend language independent ([3, 4]). However, this is not relevant to our proposal. Our goal is not to *copy* the movements but to *label* the movements in the source speech and use this characterisation as additional features to infer the intonation model in the target language.

F0 movements can be characterised using different temporal scales. On one side, global parameters can be derived from a whole sentence or paragraph (as mean F0 or range). On the other side, F0 contour can be characterised using high resolution curves. To choose the best temporal scale two aspects need to be considered. First, the scale has to be meaningful for the aspects which are going to be modelled. For instance, to label the emphasis, the use of a global parameter for the whole sentence is clearly unappropriated. The second aspect is that the chosen unit can be identified in both the source and target languages and

[1] http://www.tc-star.org

can be mapped to some extend. From this point of view, sentence parameters clearly accomplish this requirement. In most of the cases (and in most of the translation systems), one sentence in the source language is translated into one sentence in the target language. Phoneme movements are not valid because it is not easy to establish the relationship between phonemes in both languages. This problem has been extensively studied in statistical machine translation (see for instance [5]). Nowadays, most of these systems train the translation models based on single-word alignment. Of course, there are many examples where it is not possible to find one target word for each source word (as in the Spanish/English pair *Sin embargo/However*). But this can be easily solved including the *null* word in the target language [5].

In this paper we have chosen the accent group as the unit for defining F0 movements. This unit was already used in most of our previous work on intonation modelling (e.g: [6]). The accent group is defined based on the stressed words, usually, content words. In English, accent groups are defined as the stressed syllable and all preceding non-stressed syllables. In Spanish and Catalan we define accent group as the content word and all the preceding function words. This definition has been successfully used in many intonation models in Spanish (e.g.: [7]). Accent groups are a good trade-off when choosing the temporal scale. High quality contours can be modelled using piece-wise curves, one curve for each accent group, if the curves are flexible enough. At the same time, the global effect (e.g. F0 mean) can be included on each partial contour.

To extend word alignment to accent group alignment we have used a simple approach: if a content word in the source language is aligned with the target word in the target language, then associated accent groups are aligned. For instance, Figure 1 shows the word alignment of the phrase pair (*La casa blanca / The white house*). Using the word alignment shown in the figure and extending it to accent group we got (*La casa* $\Rightarrow$ *house*) and (*blanca* $\Rightarrow$ *the white*). Although this alignment is not perfect (the determinant *la* in the first Spanish accent group is shifted into the second English accent group), the main effects can be properly transferred. The most important movements are included in stressed syllables and these are properly linked. In case that a content word in one language is not aligned to one content word in the other language, the associated accent groups are not aligned and this additional feature cannot be used to derive the prosody.

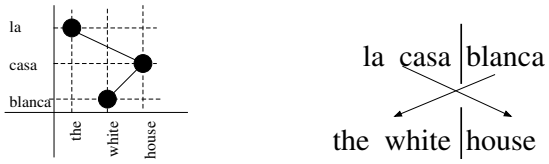

Figure 1: Word and accent group alignment.

This general framework can be used in many data-driven intonation models. In this work a superpositional model is used. The F0 contour is generated using phrase components and accent components. The phrase component spans for a intonation phrase and is modelled using a Bézier polynomial (order 4). This component models slow movements in the F0 contour. Then accent component models additional movements around the stress. In our approach we model this using also a Bézier polynomial (order 4) in the accent group. This approach is similar to the Fujisaki intonation model, but the contour is modelled using a polynomial representation instead of 2nd order filters.

For each intonation phrase a set of features derived from the text are used to generate the prosodic contour. These features are used to estimate Bézier parameters. The accent group-based contour is derived not only from the text but also from the new label that contains prosodic translation information. For this purpose, a method that allows a global estimation of the prosodic parameters (i.e. Bézier parameters) is used. This method is the Joint Extraction and Modelling Approach (JEMA) already published in [6].

Figure 2 outlines the process used to generate the prosody in the complete speech-to-speech translation system. The source speech is recognised and translated into the target language. In parallel, the source speech is analysed and annotated prosodically, as will be described in next section. Based on statistical alignment, the source accent groups are mapped to the target accent groups. The speech synthesis component includes the prosody generator modules. Usually this module derives the synthetic prosody based on features derived from text (as sentence modality, stress, number of syllables and in some cases syntactic features). In this proposal these features are extended to include the new feature derived from the source speech. A synthesiser is used to generate the speech based on the target text and the prosody.

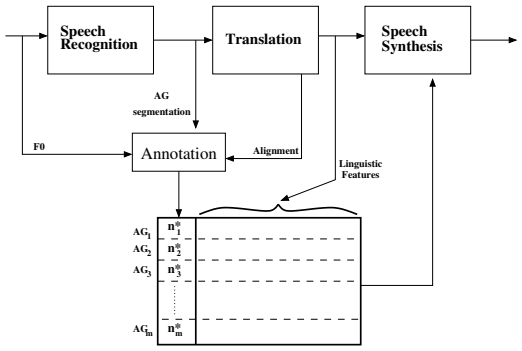

Figure 2: Annotation process uses accent group segmentation, alignment and source pitch in order to add a new feature to the vector of linguistic features for each accent group.

In order to work out relations between intonation in different languages a parallel corpus in both languages is needed. Then, this corpus is used to train the annotation model.

3. Annotation Training

The intonation clustering algorithm works on the assumption that some pitch movements of the source language have a correspondence with pitch movements of the target language. Such movements can be used as patterns that are repeated in the database. Furthermore, such patterns can be considered classes and used to code the input intonation. The coding of the input intonation is useful because it may be used as an additional input feature for the intonation module. Such classes are expected to reflect higher level linguistic information that improves the naturalness and pleasantness of the intonation.

When pitch contours are not close enough to the representative of the cluster they belong to, it is not possible to assign them to any class. Both cases are taken into account either when it happens for the source contour or the target contour. When any of them are not similar to the representative of their class, they are assigned to a NOCLASS class. Clustering is, thus, done based on source contours as well as on target contours.

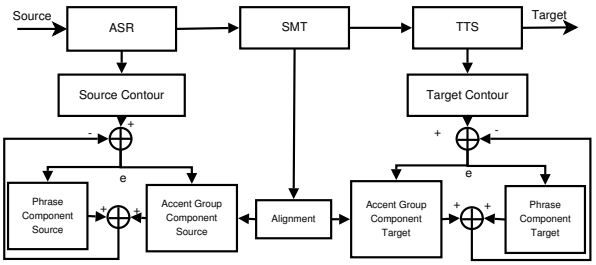

Figure 3: Illustration of the training algorithm for the prosodic model based on source prosodic information

Additionally, the automatic alignment between languages may have a word or a group of words unaligned. In this case the algorithm can not find a relationship within pitch movements and they correspond to the class named NOCLASS.

The overview of the approach is shown in Figure 4. Each class in the source language has a corresponding class in the target language. Although pitch movements in the source and target language have a relationship, they do not have to be necessarily the same, then both classes have its own intonation model. In Figure 4 we also see the use of the NOCLASS label to group pitch movements that do not have link between languages because of missing alignment information. Additionally, some links have the NOCLASS label because the pitch movement does not have a relationship between languages (i.e.: the pitch movement in the source language does not help to predict the pitch movement in the target language). In this section we propose a method to define the classes. Once this is done, the new feature is included in the target language. In the training phase, the feature is used to estimate the intonation model. During synthesis, the feature is used to derive the F0 contour.

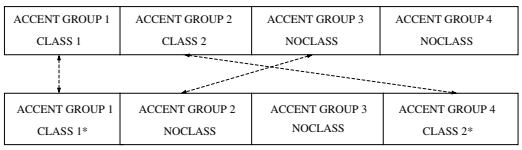

Figure 4: Example of accent group alignment and class assignment.

The goal of the intonation clustering approach is to find an arbitrary number of classes that will be patterns used to classify pitch movements in the source language. These patterns have a relationship with pitch movements in the target. Information flow through training architecture is shown in Figure 3 and steps of the algorithm are detailed below based on Figure 4.

1. **Initialisation.** It consists on assigning random classes to each accent group in the source language and label accent groups of the target language with the corresponding class according to links between source and target languages. For example, **class 1** of the source language is linked with **class 1*** of the target language, **class 2** is linked to **class 2***, and so on. In the initialisation, accent groups that do not have links between languages are assigned the class NOCLASS.

2. **Optimal patterns.** In this step the optimisation algorithm explained in [6] is used to obtain a pattern that globally approximates all the contours that belong to the class. We use a superpositional model for intonation (phrase component plus accent group). The phrase component is modeled based on linguistic features and substracted from the contour. The residual contour contains accent group information. A representative contour is built for all accent groups in the same class (i.e. representative pattern). The optimal pattern for each class is used to classify and label pitch movements of new contours, in the testing phase and in the final implementation of the speech-to-speech translation system, with intonation transfer between languages.

3. **Classification of pitch movements.** In this step the patterns obtained in the previous step are used to classify the pitch movements. Then, many pitch movements may change its assigned class because other pattern approximates it more accurately based on the root mean square error. Pitch movements are assigned to the class that minimises the approximation error both in source and target languages. In this way, we guide the clustering algorithm to find pitch movements that have a relationship between languages. NOCLASS contours are modelled using a clustering based on liguistic features.

4. **Outliers elimination.** All contours that have an Euclidean distance to the pattern higher than a predefined threshold and one standard deviation of the mean error of the class are assigned the label NOCLASS. The threshold is also used in the testing phase and in the final implementation to analyse whether a pitch movement in the source language must be assigned to a class. Not all pitch movements in the source language have a relationship with a movement in the target language.

5. **Convergence of the algorithm and addition of new classes.** The number of pitch movements that change its class are calculated. It is a measure of algorithm convergence. If the percentage of changes over the total number of pitch movements is below a threshold (in this work five percent) we consider that the algorithm almost converged and new classes can be added. Pitch movements that have a link between languages and were assigned to the class NOCLASS are assigned randomly to the added classes. Therefore, the ones that do not have a class may generate new classes. Pitch movements that belong to a class with elements below a threshold (in this work twenty) are also considered of class NOCLASS and are assigned randomly to the added classes. A class with few elements is considered that has low generalisation capabilities because there is not enough information in the database.

6. **Phrase component modelling** The phrase component of source and target is modelled using a tree as explained in [6]. After each iteration of the clustering algorithm the trees of the phrase component of source and target speaker introduce a new class. Phrase components are trained using linguistic features extracted from text of source and target languages. Therefore, accent group component is trained using acoustic features and phrase component is trained using linguistic features.

7. **Stop condition.** The clustering algorithm is considered to have converged to the necessary number of classes when the global root mean square error is not reduced in a predefined number of iterations (in our experiments we set this to ten). When the stop condition is reached,

iterations continue until the number of pitch movements that change its class is zero. Thereby, we ensure that the algorithm converged and the classes are pure with the minimum error.

4. Experiments

A parallel corpus in both languages Spanish and Catalan was used in the experiments. A bilingual person recorded it in Spanish and Catalan. In each language this person was requested to read the text in the same way. It does not mean that intonation must be the same, but that *intention* must be the same. It consists on two hundred sentences from the hotel booking domain recorded for each language. It was processed in order to train the models proposed. It was segmented into phonemes by means of the speech recognition system of UPC named RAMSES [8]. Sentences were aligned using GIZA+[5] producing links between words across sentences from both languages. In these experiments we have not taken into account neither ASR nor translation errors.

The method proposed has been compared with a baseline. It consists on the same intonation model, but without the translation models. Therefore, intonation in the target language is generated only based on the text, without the new features introduced by the prosody translation model. Then, it will be possible to analyse the improvement performed by the prosody translation model.

Both translation directions were evaluated. Results do not have to be equal in both directions. Objective measures consisted on Root Mean Square Error and correlation coefficient of the logarithm of the pitch between predicted and original contours.

4.1. Results

In Table 1 objective results are presented. In the training step a large improvement with respect to the baseline system is produced for both translation directions. This shows how the new feature based on prosody translation help to improve the modelling. The improvement is confirmed in the test set. A significant improvement in the test set shows the success of the proposed technique. However, subjective tests will give us an idea of the real improvement on naturalness.

	Train		**Test**	
Experiment	**RMSE**	ρ	**RMSE**	ρ
Baseline-Spa→Cat	0.121	0.68	0.136	0.54
Proposal-Spa→Cat	0.099	0.77	0.128	0.63
Baseline-Cat→Spa	0.113	0.69	0.135	0.56
Proposal-Cat→Spa	0.095	0.77	0.121	0.65

Table 1: Experimental results using objective measures in the train and test sets

Informal subjective tests show an improvement using the proposed feature. Therefore, both objective and perceptual results showed an improvement based on the use of the prosody translation model with respect to not using it.

5. Conclusions

In this paper we have presented a clustering algorithm which takes advantage of correlation between the intonation of the source and target speaker in the framework of speech-to-speech translation to extract patterns that can be used to improve the naturalness of text-to-speech synthesis. In many cases it is not possible to establish a set of labels to annotate the input prosody. It is valuable the use of automatic methods to extract patterns without human supervision.

These patterns are used to classify pitch movements of the input intonation. The classes are mapped from the source language onto the target language using the alignment information provided by statistical alignment. This information is an additional input feature for the intonation model of the text-to-speech synthesis module that improves naturalness because reflects semantic and pragmatic information.

Experiments show a clear improvement due to the additional information extracted from the source speaker. Objective results in the test set improved both correlation and RMSE. Intonation modelling is a field where multiple contours are possible. As a consequence, a comparison with only one pattern is a pessimistic quality measure. Furthermore, informal subjective tests show that the proposed approach has a superior MOS. As a consequence, the proposed approach has shown an improvement in the naturalness of the text-to-speech synthesis in the framework of speech-to-speech translation by the use of extracted information from the intonation of the source speaker.

In future works, and under TC-STAR project, a larger expressive parallel corpus will be recorded with a parliamentary style. This database will be used to test the presented technique in other languages such as English-Spanish. Four bilingual speakers will take part in the recordings. A parallel corpus of two languages and four speakers will be available for spoken translation studies.

6. References

[1] S. Werner and E. Keller, *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*. E. Keller, 1994, ch. Prosodic aspects of speech, pp. 23–40, Chichester. John Wiley.

[2] N. Campbell, "Speech & expression; the value of a longitudinal corpus," in *Proceedings of LREC*, 2004, Lisbonne, Portugal.

[3] D. Hirst and A. D. Cristo, Eds., *Intonation Systems: A Review of twenty languages*. Cambridge University Press, 1998.

[4] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *EUROSPEECH 2001*, September 2001, Aalborg, Denmark.

[5] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proc. of Association for Computational Linguistics*, Hongkong, China, October 2000, pp. 440–447.

[6] P. D. Agüero and A. Bonafonte, "Intonation modeling for TTS using a Joint Extraction and Prediction Approach," *Proceedings of the International Workshop on Speech Synthesis*, 2004.

[7] D. Escudero, "Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en Español." *PhD Thesis, Universidad de Valladolid*, 2002.

[8] A. Bonafonte, J. B. Mariño, A. Nogueiras, and J. A. R. Fonollosa, "RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC," in *VIII Jornadas de Telecom I+D (TELECOM I+D'98), Madrid,Spain*, October 1998.