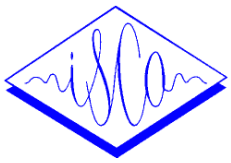

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PROTRAN: A PROSODY TRANSPLANTATION TOOL FOR TEXT-TO-SPEECH APPLICATIONS

B. Van Coile (1,2), *L. Van Tichelen* (1), *A. Vorstermans* (1,2), *J. W. Jang* (1), *M. Staessen* (1)

(1) Lernout & Hauspie Speech Products, Ieper, Belgium
(2) ELIS, University of Gent, Belgium

ABSTRACT

This paper describes the technique of *Prosody Transplantation*, its advantages and disadvantages. Special attention is paid to a development tool, called *ProTran*. This integrated tool running under MS-Windows was designed to automate and speed-up all steps that are involved in the prosody transplantation process.

PROSODY TRANSPLANTATION

Introduction

Speech quality of text-to-speech has reached a level that is sufficient for a lot of applications. However, none of the available text-to-speech systems can fully replace a human reading aloud. Deficiencies in synthetic prosody are undoubtedly partly responsible for this quality difference between synthetic and natural speech. In this paper we describe the concept of Prosody Transplantation as a method to improve the quality of synthetic speech. One can take advantage of this technique in applications where at least some of the messages to be synthesized are fixed and known in advance. This is for example the case in interactive voice response systems or in announcement systems such as automatic traffic messaging.

The concept

The technique of Prosody Transplantation is based on the idea of transplanting (copying) intonation and duration values from a recorded *donor* message (human speech) to the phonetic transcription of the same message. The *enriched phonetic transcription* (EPT) thus obtained, can then be used as input for our TTS system, by-passing the linguistic and prosodic modules. Only the segmental synthesis module and the synthesizer module are used. The output of this Phonetics-To-Speech (PTS) synthesis is high quality synthetic speech.

The prosody transplantation process

The prosody transplantation process includes several steps. A recording of the message is made with the selected prosody donor. After speech analysis, the speech signal is segmented and labeled. Phoneme durations are derived from this segmentation and labeling. The natural pitch contour is analyzed and approximated in the log-domain by a perceptually equivalent piece-wise linear contour. Finally, the phonetic representation, the phoneme durations and the breakpoints of the intonation contour are combined into an enriched phonetic transcription. The following example shows such an EPT (textual representation) for the English sentence *Thank you for your attention.*

#T[104]æ[74(0,98)]N[47]k[107(10,81)]j[14(0,106)]
u[44]f[93(0,91)]o[47(0,102)]r[29]j[68(0,98)(30,90)]
o[50(0,96)]r[71] $[45(0,93)]-t[108]E[70(0,102)]
n[68]-S[96]$[56]n[106(30,83)(100,83)]#

The first value between square brackets is the phoneme duration (in ms), optionally followed by one or more intonation breakpoints. Each breakpoint consists of a location value (in ms) relative to the beginning of the phoneme, followed by a pitch value (in ST/4; reference 50 Hz).

Advantages

The main advantage of Transplanted Prosody is the improvement in speech quality compared to full text-to-speech. A substantial quality increase is obtained while the bitrate needed to store EPT's is still very low. In our current implementation of Transplanted Prosody, we store an EPT with a bitrate of less than 300 bit per second. This is sufficient to synthesize high quality synthetic speech. This means that PTS in combination with Transplanted Prosody offers an interesting alternative to speech coding in applications with a large amount of fixed messages .

Synthetic speech with Transplanted Prosody inherits the segmental characteristics (speech timbre) from the speech synthesis system that is used. Our text-to-speech system is based on a segment concatenation technique (diphones and a minority of larger units). Therefore, our transplanted prosody speech has the timbre of the original voice that was used for the development of the segment data base.

However, the prosodic features of the synthetic speech are inherited from the speech recorded with the prosody donor. This means that, in contrast with classic speech coding techniques, only the prosody and not the timbre of the speaker (the prosody donor) is retained. This is a clear advantage for applications with a very large message set. If speech coding is used, the availability of the original speaker is very important for maintenance and updates of the message set. However, the long-term availability of a speaker can never be guaranteed. With the concept of Transplanted Prosody, this speaker problem can be overcome. Although a speaker is still needed to record all the messages, speech timbre does not depend on the selected speaker anymore. This means that even with a new prosody donor, the message set can easily be updated.

One of the most important features of Transplanted Prosody and PTS is the possibility to combine these techniques with full text-to-speech in an integrated way. This is very interesting for applications where fixed messages are combined with variable information. Text-to-speech can be used for the variable parts, while PTS and Transplanted Prosody are used for the fixed parts. This flexibility is especially important for applications that use a number of fixed *carrier phrases* with *open slots* such as:

You have ordered [number] [article name].
Last call for flight [flightnumber] with destination [city], please proceed to gate [number].

The open slots are filled at run-time with variable, often orthographic information. As the carrier phrases are fixed and known in advance, they can be stored as EPT. During the development of the application, all carrier phrases are recorded at least once with their open slots filled in with typical slot fillers. From these recordings, EPT's are extracted for the carrier phrases. An important aspect is the synthesis of prosody for the open slots. We are currently using the following approach. Phoneme durations for the slot-fillers are calculated by means of the standard duration model of the text-to-speech system. This model uses features that are normally extracted from the input text at run-time. As the carrier phrases are stored as EPT's, some of the features cannot be derived at run-time anymore. Instead, we determine them off-line (i.e. during the development of the application). They are then stored in the application as part of the slot description. Intonation contours for the open slots are calculated by means of slot specific, small intonation models. These models are derived from the standard intonation model. Their main purpose is to enable only those standard contours that are compatible with the stored contour of the corresponding carrier phrase.

Drawbacks

Although the transplanted prosody concept is interesting for a lot of applications, there are also some important drawbacks. First of all, the method is only useful for those applications where at least part of the message set is fixed. Moreover, speech recordings are needed for the prosody transplantation process. If this process is done manually, it requires a lot of expertise and is very time consuming. This means that the usefulness of Prosody Transplantation really depends on the availability of tools that facilitate and speed up the transplantation process significantly.

PROTRAN: A PROSODY TRANSPLANTATION TOOL

The ProTran tool has been designed to automate and speed-up the prosody transplantation process. It has been implemented under MS-Windows as an interactive tool. The central part of the system consists of the ProTran Manager and the Graphical User Interface (GUI). They can be considered as a framework in which different modules (DLL's) can be plugged in. This means that we can easily replace one algorithm by another algorithm as long as the functionality and the interface of the different modules comply to our ProTran specification. Figure 1 gives a schematic representation of this approach.

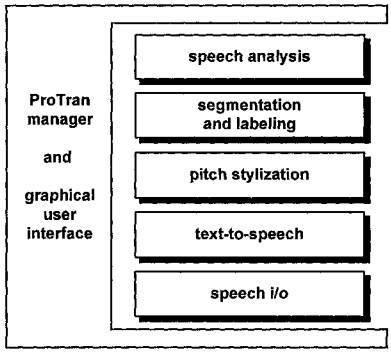

Figure 1: *The ProTran tool consists of a framework in which different modules can be plugged in*

The tool supports all steps of the prosody transplantation process. This includes: speech recording and text input, grapheme-to-phoneme conversion, speech analysis, pitch stylization, segmentation and labeling, EPT creation and EPT optimization.

Speech recording and text input. Prosody Transplantation for one message starts from the orthographic representation and the corresponding acoustic realization. Text input as well as speech input can be read from file or can be entered by means of keyboard and microphone.

Speech analysis. Depending on the segmentation and labeling method used, the speech signal is analyzed with the appropriate analysis algorithm. For the automatic segmentation and labeling system described further, speech is analyzed with an auditory model. The output of this module consists of an auditory spectrum, a voicing evidence, a pitch frequency and 5

samples of an intensity function every 10 ms. A more detailed description of the auditory model can be found in [3].

Grapheme-to-phoneme conversion. The automatic segmentation and labeling system needs a.o. the phonetic representation of the message. This phonetic transcription is obtained automatically by calling the grapheme-to-phoneme routines from the TTS DLL (or by entering the transcription manually). The GUI supports the editing of the transcription.

Segmentation and labeling. Both manual and automatic segmentation and labeling are supported by the ProTran tool. Normally the automatic segmentation and labeling algorithm is used first. If needed, the results can be validated and modified manually. The system also supports the interaction between manual and automatic segmentation and labeling. This means that manually specified and labeled boundaries can be used as constraints for the automatic segmentation and labeling. This offers a powerful and flexible method for correcting segmentation and labeling errors.

Pitch Stylization. Starting from the pitch and voiced/unvoiced information coming from the speech analysis module, a piece-wise linear (PWL) approximation of the intonation contour is created. Both manual and automatic stylization are supported. The GUI allows the user to add, delete or modify intonation breakpoints. Manually specified breakpoints can also be used by the automatic method as extra constraints for the stylization process.

EPT creation. The phoneme durations are derived from the segmentation and labeling results. The phonetic representation, the phoneme durations and the breakpoints of the stylized intonation contour are combined into an enriched phonetic transcription.

EPT optimization. The enriched phonetic transcriptions can be synthesized using the TTS module. The GUI allows the user to change phoneme durations, to modify the stylized intonation contour and to listen to the synthesized speech interactively.

Some of the prosody transplantation steps just described rely on language specific modules. Up to now, the following languages are supported: (American) English, French, German, Spanish and Dutch [2,4].

In the following paragraphs we will focus on two algorithms, one for automatic pitch stylization and one for automatic segmentation and labeling.

Automatic segmentation and labeling

The automatic segmentation and labeling system is shown in figure 2. It can be described as a segment-based Dynamic Programming/Multi-Layer Perceptron system. The input of the system consists of speech parameters from the auditory model, the phonetic representation and optionally some phone boundaries and labels that were already validated or determined manually. An initial segmentation is obtained by generating landmarks at times of maximum acoustic variation. Only those landmarks which have a high probability of being phonetic segment boundaries are retained. This probability is estimated by means of a multi-layer perceptron (MLP). Starting from the initial segmentation, different candidate phonetic segments can be constructed by merging initial segments. In the phonetic segment classification module, a phonetic/non-phonetic classifier (MLP) is used to estimate the probability that a candidate segment coincides with a true phonetic unit (phone). The segment candidates also receive broad phonetic class evidences using an MLP that distinguishes between 5 different classes: vowel, sonorant, fricative, plosive and closure/silence.

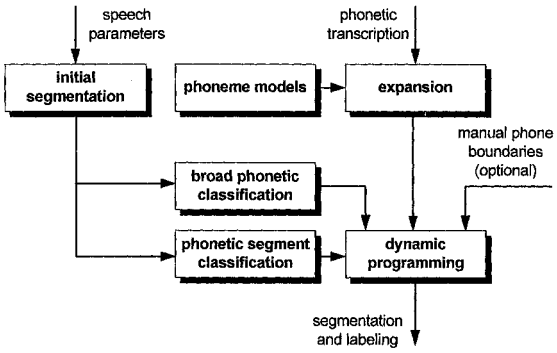

Figure 2: *Outline of the automatic segmentation and labeling system.*

The system also uses simple statistical phoneme models describing the phone component structure of the phonemes in terms of broad phonetic classes. By concatenating phoneme models, the phonetic transcription of the utterance is converted into a statistical utterance model. A dynamic programming technique is then used to determine the best segmentation and labeling given the utterance model. During this optimization process the manually specified phone boundaries and labels are taken into account as extra optimization constraints.

The segmentation and labeling system was originally trained on a multi-speaker data base of hand-labeled Dutch (Flemish) utterances (20 speakers; 13 sentences each). However, this baseline system can be adapted to new languages and/or new speakers without requiring manually segmented and labeled utterances. If a new language is involved, the phoneme list and the corresponding phoneme models (maximum length phone sequences described in terms of broad phonetic classes) must be specified. In a first step, the new data are automatically segmented and labeled using the baseline system. These alignments then provide the observations required to update the broad phonetic classification network. The other MLP's are left unaltered, as they were found to be speaker and language independent. Finally, the transition

probabilities of the statistical phoneme models are derived from alignments obtained with the already updated classification network.

The method has already been formally evaluated on several languages. More details on the segmentation and labeling method as well as evaluation results can be found in [4].

Pitch stylization

Starting from the pitch and voiced/unvoiced information, the pitch stylization system generates a piece-wise linear (PWL) approximation of the intonation contour (in the log-domain). If available, the segmentation and labeling results are also used as additional information to guide this process. Optional manual breakpoints are used as extra constraints during the stylization process which is performed in two steps. In the preprocessing module, a continuous pitch contour is created for the complete message. The original contour is split into voiced intervals of continuous pitch. By examining the sequence of these intervals, corrections and modifications are made to the original contour. The system also uses a strategy to extend the contour to those parts labeled as unvoiced. As a result of the first module, a continuous intonation contour is obtained with a reduced number of pitch errors, creaky voice effects and micro-prosody influences.

The second module of the system performs the actual PWL approximation. The system uses an iterative procedure. At each iteration step, a breakpoint is added by looking at a weighted error function (i.e. a weighted difference between original and stylized contour). The weighting function is derived from the error function, taking into account the number of breakpoints already assigned, and therefore becomes more detailed as the number of breakpoints increases. As a consequence, the first breakpoints will describe the more global pitch trends. As more breakpoints are added, the algorithm starts focusing on more local pitch movements. Two criteria are used to stop this iterative process: a threshold criterium using a distance measure and/or a limit on the number of breakpoints.

Figure 3 shows a stylization example for for a short Spanish sentence.

No elaborate formal evaluation of the method has been performed yet. However, the method has already been used for the development of several transplanted prosody applications. During some of these developments, we have used a distance threshold that results in the same average number of breakpoints as in manual stylizations, obtained from one of our team members who is an expert intonologist trained in the IPO stylization approach [1]. The automatic stylization results were validated by the same expert. Less than 10 % corrections (insertions, deletions and modifications) to the breakpoints were needed to optimize the PTS output. This number can be reduced by lowering the stop threshold at the cost of more breakpoints and a somewhat higher EPT bitrate.

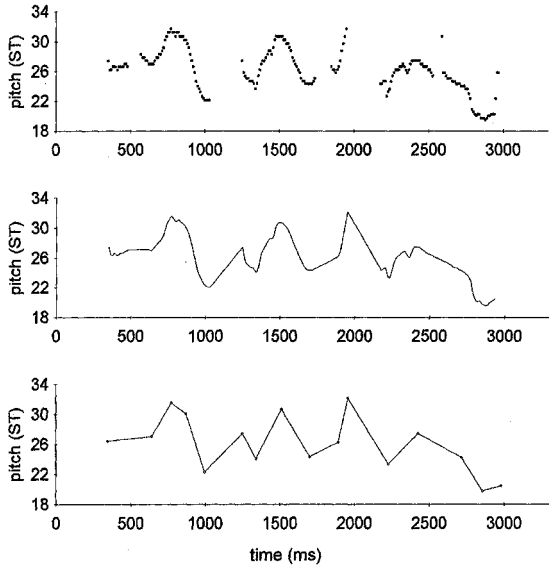

***Figure 3**: Pitch stylization for the Spanish sentence 'Los alumnos (que viven lejos) llegan tarde'. Original pitch contour (top), output of the preprocessing (middle) and stylized contour (bottom).*

SUMMARY

Prosody Transplantation is an interesting method to improve the speech quality in text-to-speech applications where at least some messages (or parts of messages) are fixed and known beforehand. Thanks to the availability of a flexible tool, called *ProTran*, the prosody transplantation process is no longer a difficult and time-consuming task.

REFERENCES

[1] R. Collier (1989), "Intonation Analysis: the Perception of Speech Melody in Relation to Acoustics and Production," Eurospeech-89, pp. 38-44.

[2] B. Van Coile (1993), "On the Development of Pronunciation Rules for Text-to-Speech Synthesis," Proc. Eurospeech 93, vol. 2, pp.1455-1458.

[3] L.M. Van Immerseel and J. P. Martens (1992), "Pitch and voiced/unvoiced determination with an auditory model," JASA vol. 91 no. 6, pp. 3511-3526.

[4] A. Vorstermans and J.P.Martens (1994), "Automatic Labeling of Speech Synthesis Corpora," ICSLP-94.