Prosody transplantation using unit-selection: Principles and early results

Mythri Thippareddy and V. Ramasubramanian
PES Institute of Technology - Bangalore South Campus (PESIT-BSC)
Bangalore, 560100, India
mythri.thippareddy@gmail.com, v.ramasubramanian@pes.edu

***Abstract*—We address the problem of prosody transplantation or conversion in TTS in a 2-stage unit-selection framework, which uses the units determined in a 1st-stage conventional unit-selection to be mapped to units in a 2nd-stage prosodic unit-selection. This retrieves units from the 2nd-stage unit-database that has an associated prosody in the prosodic-style of the 2nd-stage unit-database and allows further incorporation of the prosodic information from the 2nd-stage unit-selection on to the spectral information of the 1st-stage unit-selection in any of the conventional synthesizer system. We introduce a Viterbi algorithm for optimal unit-selection on pre-grouped units and identify several experimental scenarios that will impact the nature of the prosody that can be transplanted from the 2nd-stage unit-selection. Preliminary results confirm the validity of the framework, and show the promise of such a prosody transplantation scheme.**

I. Introduction

Text-to-speech synthesis (TTS) systems presently provide speech of such high quality as considered adequate for most applications to date, such as spoken dialog systems, car navigation, mobile applications, talking book etc. While this assumes that such TTS output have adequate naturalness considered acceptable, there still remains a problem in TTS speech quality that is currently considered difficult, not yet solved adequately, and presenting a high degree of challenge - this is the problem of incorporation of prosody in the synthesized speech, as can be derived only from the input text. The terms prosody here refers to a broad range of diverse aspects that the synthesized speech can have, such as naturalness at one end, to expressive or emotional speech at the other end, with varying degrees of prosody in between, dictated by the specific task in question where the TTS is functional, such as for example, spoken dialog systems, where certain kinds of interaction with the user may require special prosody (stressed speech of some part of the retrieved information, query clarification, etc.). Apart from such application driven need for appropriate prosody control in TTS, the problem of general expressive prosody, with a certain extent of control on the degree and style of prosody still remains a hard problem with high academic value and importance [1], [2].

The most direct approach to incorporating prosody in a TTS for a given input text is to simply 'predict' the prosody from the input text. This has attracted considerable attention (see for example a review and recent approaches in [3]) and has solutions ranging from rule-based approaches (that use syntactic and semantic structures and information in the text to arrive at an appropriate prosody prediction) to data-driven machine-learning methods such as CART or neural-network based functional mappings.

In this paper, we approach the problem of prosody incorporation into the synthesized speech along the lines explored earlier by Prudon et al. [4]. Here, the prosody was obtained by 'selection', i.e., unit-selection from a specific prosody-database, for an input text. The selected prosodic parameters, primarily the duration of units, intonation and gain contours, are combined with the phonetic information from the text (derived via G2P) and synthesized by a diphone synthesis system using MBROLA. By this, the authors completely circumvented syntactic analysis and syntactic-prosody rules, and used the prosodic part of the corpus obtained by such a unit selection.

Specifically, taking cue from this method, wherein the prosodic part of a unit-selection of units corresponding to the input units seem to perform an acceptable prosody 'transplantation', we propose a 2-stage unit-selection for acquiring an appropriate prosody of the input text, where the first stage performs a conventional unit-selection from a 1st-stage unit-database to derive the units that are to be synthesized. These units, in symbolic form are used in the 2nd-stage unit-selection using a 2nd stage unit-database, only to retrieve units that match the units of the 1st-stage output but have an associated prosody as present in the 2nd stage unit-database. This prosody (mainly the duration of units, pitch contour and gain contour) are combined with the phonemic part of the 1st-stage output and synthesized by any conventional synthesizer (e.g. PSOLA, MBROLA, or LPC) to yield synthesized speech that has the speaker characteristics of the 1st-stage unit-database, but the prosody style of the 2nd-stage unit-database.

We first present the basic principles and algorithmic aspects of this 2-stage unit-selection framework for the primary purpose of incorporating a desired style of prosody in the output synthesized speech, and follow it with a brief outline of experimental scenarios that such a framework entails and some early results of such a framework.

II. Two stage unit-selection prosody transplantation

Fig. 1 shows a schematic of the 2-stage unit-selection framework for prosody transplantation we are proposing here.

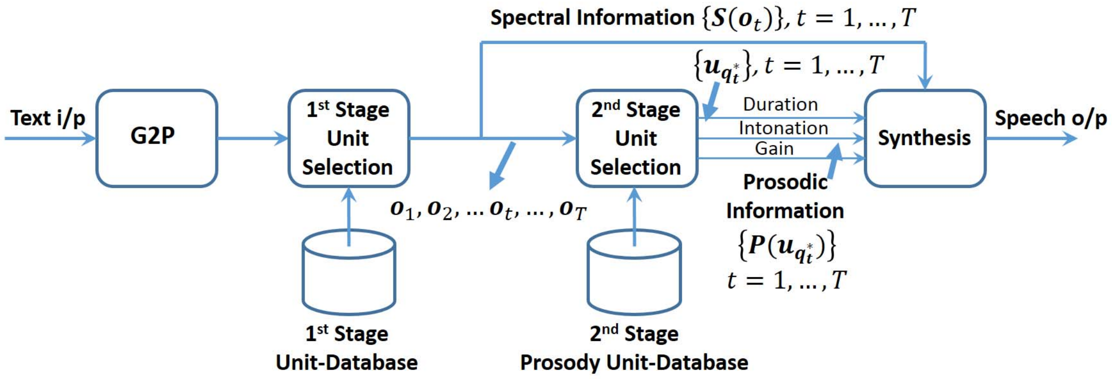


Fig. 1. *Schematic of proposed 2-stage unit-selection framework for prosody transplantation*

In this framework, the 1st-stage unit-selection is the conventional part of a concatenative TTS. It is driven by the output of the G2P which generates the target unit sequence vectors from the given input text. The 1st-stage unit-selection then selects appropriate units from the 1st-stage unit-database, which can be used to drive a concatenative synthesizer such as LPC, PSOLA or MBROLA. However, the system does not have a synthesizer in the 1st-stage. Instead, the spectral component of the 1st-stage units selected from the 1st-stage unit-database are sent directly to a synthesizer at the 2nd-stage, for a subsequent combination with the prosodic information derived from the 2nd-stage unit-selection. The 2nd-stage unit-database carries the desired prosody and could be spoken by the same or different speaker as the 1st-stage unit-database. For instance, the 2nd-stage unit-database could be in different styles as desired in the synthesized speech of the overall system, such as, news-reading style, story-telling styles, or various emotional styles etc. It is important to note that the 1st-stage unit-database is of a typical size as is required for high qualilty synthesis, while the 2nd-stage unit-selection could be smaller, being able to adequately provide the prosodic-stylistic aspects of a given unit-sequence. This is more or less the prime motivation for the prosody transplantation in the earlier work [4] which forms the basis of the work we report here.

Note that it is necessary to derive the spectral part for later combination with prosodic parameters only in the case of LPC type of source-filter synthesizer; instead, if the synthesizer in the 2nd-stage is PSOLA, the 1st-stage unit-selection output units can be retained as waveform segments which are further modified in duration and pitch by means of time-scale and pitch-scale modifications as is commonly done in STRAIGHT frameworks. This is a relatively minor detail, but assumes importance when we consider the fact that the synthesized speech continues to carry the speaker characteristics of the 1st-stage unit-database, but with the prosodic style associated with the 2nd-stage unit-database, spoken by the same or different speaker as the 1st-stage unit-database.

To help describe the Viterbi algorithm we use in the 2nd-stage unit-selection, we first denote the unit-sequence output of the 1st-stage unit-selection as $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T$. Each of these could be vectors with components indicating the primary phonetic (or any unit) label, left-context, right-context, duration, pitch, gain or any other features that characterize the unit selected from the 1st-stage unit-database, and that is considered relevant for the 2nd-stage prosody-selection. Note that a vector $\mathbf{o}_t$ in this sequence is derived from the 1st-stage unit-database, and not from the input text, and hence can afford to have features derived by signal-processing on the 1st-stage unit-database. Of these, the question of which of these are relevant for the 2nd-stage prosody-selection is left to be considered under ‘experimental scenarios’ outlined in Sec. III.

Let the 2nd-stage unit-database be represented as a sequence of N units, $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_N)$. In the grouping based Viterbi, we first form groups $G_1, G_2, \ldots, G_t, \ldots, G_T$, with group G_t corresponding to input vector $\mathbf{o}_t$, and consisting of all units in the unit-database $\mathcal{U}$ that share some feature components, i.e. for instance, if an input vector $\mathbf{o}_t$’s primary label is a particular phoneme, the group G_t will have all units in $\mathcal{U}$ whose primary label is also the same phoneme, i.e., the group G_t is simply a collection of all phonemic units from the unit-database whose primary labels are same as that of the input vector $\mathbf{o}_t$. The need to define this in a general sense, is to allow scope for a flexible definition of the group by way of allowing various criteria such as is outlined in the section on ‘Experimental Scenarios’. Fig. 2 shows the trellis formed by the above defintions of the input unit sequence $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T$ in the x-axis and the corresponding groups $G_1, G_2, \ldots, G_t, \ldots, G_T$ in the y-axis.

The 2nd-stage unit-selection performs a grouping-based unit-selection on this trellis, as outlined in the algorithm shown in Fig. 3. Let i denote the index of the current unit $\mathbf{u}_i$ (in group G_t) being analyzed for the input unit $\mathbf{o}_t$, and let j be the index which spans all the units in group G_{t-1} considered for the previous input unit $\mathbf{o}_{t-1}$. Let the target cost between $\mathbf{o}_t$ and $\mathbf{u}_i$ be denoted by $d_u(t, i)$. Let the concatenation cost defined between two units $\mathbf{u}_j$ and $\mathbf{u}_i$ in the unit database be donated by $d_c(j, i)$. Let $D(t, i)$ denote the accumulated cost of the best path reaching the co-ordinate (t, i) in the trellis, and let $\psi(t, i)$ record the unit in G_{t-1} at $t - 1$ that is part of the best path reaching (t, i).

$d_c(j, i)$ is the concatenation costs that takes on a value of

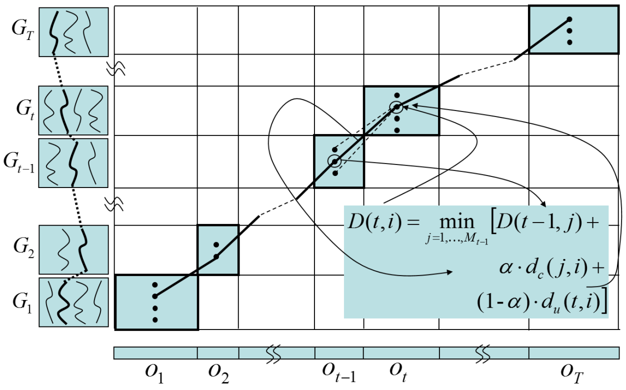
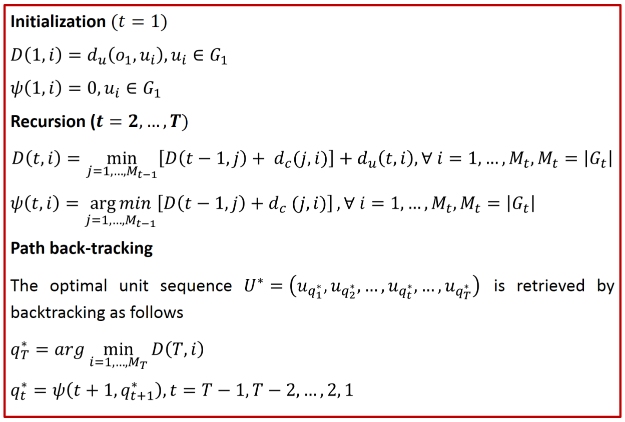

G_T
G_t
G_{t-1}
G_2
G_1
$D(t,i) = \min_{j=1,\dots,M_{t-1}} \left[D(t-1,j) + \alpha \cdot d_c(j,i) + (1-\alpha)\cdot d_u(t,i)\right]$
o_1
o_2
o_{t-1}
o_t
o_T

Fig. 2. *Trellis on which the Viterbi algorithm finds the optimal unit-selection from grouped units*



```
**Initialization** ($t = 1$)
$D(1,i) = d_u(o_1, u_i), u_i \in G_1$
$\psi(1,i) = 0, u_i \in G_1$
**Recursion** ($t = 2, \dots, T$)
$D(t,i) = \min_{j=1,\dots,M_{t-1}} [D(t-1,j) + d_c(j,i)] + d_u(t,i), \forall\, i = 1,\dots,M_t, M_t = |G_t|$
$\psi(t,i) = \arg\min_{j=1,\dots,M_{t-1}} [D(t-1,j) + d_c\,(j,i)], \forall\, i = 1,\dots,M_t, M_t = |G_t|$
**Path back-tracking**
The optimal unit sequence $U^* = (u_{q_1^*}, u_{q_2^*}, \dots, u_{q_t^*}, \dots, u_{q_T^*})$ is retrieved by backtracking as follows
$q_T^* = arg \min_{i=1,\dots,M_T} D(T,i)$
$q_t^* = \psi(t+1, q_{t+1}^*), t = T-1, T-2, \dots, 2, 1$
```

Fig. 3. *Viterbi algorithm for finding optimal unit-selection from grouped units*

0 if u_j and u_i are contiguous in the 2nd-stage unit database, and takes on a value of the Euclidean distortion between the last frame of the acoustic segment annotated as u_j and the first frame of the acoustic segment annotated as u_i, if they are not contiguous.

The above algorithm corresponds to the trellis shown in Fig. 2. In the above algorithm, selection of optimal units are based on equal weights assigned to target cost and concatenation cost. A more general unit-selection uses weighting of these two costs, so as to control the degree of unit-matching against the degree of contiguous units that can be selected reflecting naturally occuring consecutive units in the unit-database, with intrinsically high degree of natural co-articulation. This calls for modification of the recursions in Fig. 3; let α be the weight assigned to the concatenation cost; then the modified equation for $D(t,i)$ is given as

$$D(t,i) = \min_{j=1\dots,M_{t-1}} \{D(t-1,j) + \alpha \cdot d_c(j,i)\} + (1-\alpha)\cdot d_u(t,i) \tag{1}$$

The above algorithm yields the optimal unit sequence $U^* = (\mathbf{u}_{q_1^*}, \mathbf{u}_{q_2^*}, \dots, \mathbf{u}_{q_t^*}, \dots, \mathbf{u}_{q_T^*})$ from the 2nd-stage corresponding to the input unit-sequence $\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t, \dots, \mathbf{o}_T$ from the 1st-stage unit-selection, under the constraints imposed by α that trades unit-cost and concatenation cost. For any given α, the retrieved unit sequence U^* represents the units in the 2nd unit-database that are the best match to the input unit sequence in terms of the feature vectors that define $\mathbf{o}_t$ and $\mathbf{u}_{q_t^*}$ (in addition to the concatenation cost constraints between the chosen units in U^*). Therefore, once U^* is identified, it ensures that the phonetic content of the 1st-stage output is satisfactorily retrieved; given that the 2nd unit-database has a distinct prosodic style and based on the premise and finding in [4] about prosodic transplantation, it can be expected that the prosody associated with U^* will be typical of the style of the 2nd-stage unit-database, and could be transplanted on to the phonetic content of the unit sequence output of 1st-stage unit-selection, i.e., $\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t, \dots, \mathbf{o}_T$.

If $S(\mathbf{o}_t)$ denotes the spectral part of $\mathbf{o}_t$ (and hence the phonemic identity of unit $\mathbf{o}_t$) and $P(\mathbf{u}_{q_t^*})$ denotes the prosodic part of $\mathbf{u}_{q_t^*}$, then the synthesizer at the 2nd-stage can transplant the prosodic-style of the 2nd-stage unit-database onto the phonetic information of the input text, by synthesizing speech as a combination of $S(\mathbf{o}_t)$ and $P(\mathbf{u}_{q_t^*})$. This is as shown in Fig. 1. For instance, in a LPC synthesizer, $S(\mathbf{o}_t)$ could be the prediction co-efficients of frames in the unit $\mathbf{o}_t$ and $P(\mathbf{u}_{q_t^*})$ would be the set of parameters (duration of unit $\mathbf{u}_{q_t^*}$, pitch and gain contours of frames in unit $\mathbf{u}_{q_t^*}$), and the LPC synthesizer would perform a frame-by-frame synthesis of the unit ($\mathbf{o}_t$ with the prosody of unit $\mathbf{u}_{q_t^*}$, after 'duration-modification' of $\mathbf{o}_t$ to have the same duration (i.e. number of frames) as $\mathbf{u}_{q_t^*}$. Alternately, as indicated earlier, $S(\mathbf{o}_t)$ could represent the speech waveform of unit $\mathbf{o}_t$ and $P(\mathbf{u}_{q_t^*})$ the set of parameters (duration of unit $\mathbf{u}_{q_t^*}$ and its average pitch), which in turn can be used in a TD-PSOLA or STRAIGHT kind of technique that can perform a time-scale and pitch-scale modification to perform the desired prosody transplantation.

III. Experimental scenarios

In this section, we outline some scenarios that present itself in performing a prosody transplantation as above:

1) The weighing factor α trading off unit cost against concatenation cost controls the degree of match between the primary phonetic identity between the units $\mathbf{o}_t$ and $\mathbf{u}_{q_t^*}$. Values of α close to 0 cause fragmentation of the unit sequence U^*, while values of α close to 1 cause highly contiguous groups of units in U^*. This in turn can be expected to result in perceptually acceptable contiguous and naturally continuous pitch and gain contours (and durations) of the prosodic information from the 2nd-stage unit-database.
2) The degree of match between $\mathbf{o}_t$ and any unit $\mathbf{u}_{q_t} \in G_t$ is determined by the nature of features comprising these vectors. These features can be fine-grained, such as the actual fine-category phoneme class, or coarse-grained, such as various levels of broad-category phoneme class, and this in turn controls the kind of prosodic pattern that is retrieved based on the underlying match between the unit vectors thus defined, at fine- or coarse-grained manner. At one extreme, if the coarse-grained representation

is of the form of only consonant / vowel kind of categorization, it can be expected that the associated prosody will correspondingly reflect a coarse approximation to the broad-category sounds of the text in question.

3) Note also that the very definition of the group corresponding to unit $\mathbf{o}_t$ is a kind of pre-processi taking into account the question of at what granular the grouping should be done. To this extent, the groupi by itself also determines the above effect of prosc transplantation.
4) The kind of feature components that define $\mathbf{o}_t$ and a unit $\mathbf{u}_{q_t} \in G_t$ also control the kind of prosody tl can be expected to be transplanted. For instance, addi duration, intensity and pitch as feature components to unit vectors and combined with high weighting of ur... cost (small α) can result in a close to exact copy of the prosody latent in the 1st-stage unit-database by the 2nd-stage prosody; this is a redundant process, since then, the 2nd-stage adds no new information to the latent prosody of the 1st-stage unit-selection. On the contrary, not specifying the prosodic parameters (duration, intensity and pitch) and/or having a high α (i.e., low weighting of unit cost), can offer a high degree of freedom for the 2nd-stage prosody to be associated with the underlying phonemic identity in the units $\mathbf{o}_t$ and any unit $\mathbf{u}_{q_t} \in G_t$.

Thus, from the above considerations, it can be seen that the proposed framework of a 2-stage unit-selection (to find and incorporate an appropriate prosody transplantation from the 2nd-stage unit-database) provides considerable flexibility and possibilities in controlling the nature of the prosody.

IV. Early results

While much of the above scenarios need to be ascertained and assessed for their perceptual acceptability, we have carried out baseline experiments to reveal the underlying potential and functionality of the proposed framework. We report here a basic experiment by way of controlling α (i.e. Item 1 in the preceding section) in producing a good prosody transplantation and perceptually intelligible synthesis and distinct prosody style in the synthesized speech.

For this, we have performed all experiments on an Indian language 'Kannada' using the Festival 1st-stage speech synthesis followed by the proposed 2nd-stage Viterbi as discussed here. The 1st-stage unit-database has 600 sentences (in mostly neutral or close to news-reading prosody style) and the 2nd-stage unit-database has 300 sentences in 'story-telling' prosody style.

As a first scenario, we show the behavior of the 2nd-stage unit-selection Viterbi to select highly contiguous groups of longer units, when α varies from 0 to 1. Fig. 4 shows the number of contiguous groups of various number of units (from 1 to 10) for $\alpha = 0$ to 1. It can be seen that while $\alpha = 0$ causes the most fragmentation of the units, i.e., large number of blocks of size 1, with increasing α decreasing the number of blocks of size 1, and increasing the number of longer contiguous blocks. However, this trend quickly saturates for $\alpha = 0.2$ and remains more or less stable until $\alpha = 1$. This can also be noted from Fig. 5 which shows the average number of units per contiguous block as a function of α.

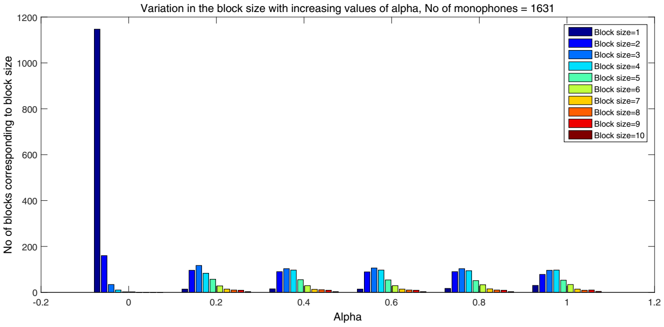


Fig. 4. *Effect of α on the 2nd stage unit-selection*

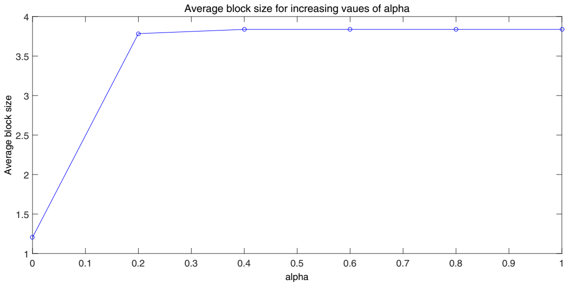


Fig. 5. *Effect of α on the average number of units per contiguous block*

As part of the same experimental scenario of varying α, we synthesize speech from the 2nd-stage unit-selection directly by concatenating these units, without combining it with the spectral segments of the 1st-stage unit-selection output and without any further prosodic modification, i.e., the unit-segments $U^* = (u_{q_1^*}, u_{q_2^*}, \ldots, u_{q_t^*}, \ldots, u_{q_T^*})$ are synthesized 'as is'. The results of this are shown in Fig. 6. Some of the observations from this figure are: i) the 1st-stage unit-selection output (2nd plot from top) closely resembles the reference speech, as is expected from a good unit-selection synthesis such as Festival, ii) the 3rd panel from top shows the speech synthesized by the 2nd-stage unit-selection for $\alpha = 0$. It can be seen that since the concatenation cost is weighted by 0, it plays no role at all, and the Viterbi emphasizes unit-cost in the cost minimization and yields a highly fragmented unit sequence; this is seen as a 'patchy spectrogram' (indicating severe unit-to-unit spectral discontinuities), marked as vertical solid lines, iii) the bottom panel shows the speech synthesized by the 2nd-stage unit-selection for $\alpha = 1$. It can be seen that since this results in giving 0 weight to unit-cost (and hence high weightage to concatenation cost), the emphasis of the Viterbi algorithm is to find groups of units with high contiguity; this can be noted from the highly natural contiguous groups, marked with solid lines indicating long contiguous group boundaries, within which are the actual units with natural continuity to adjoining units, marked by dashed vertical lines.

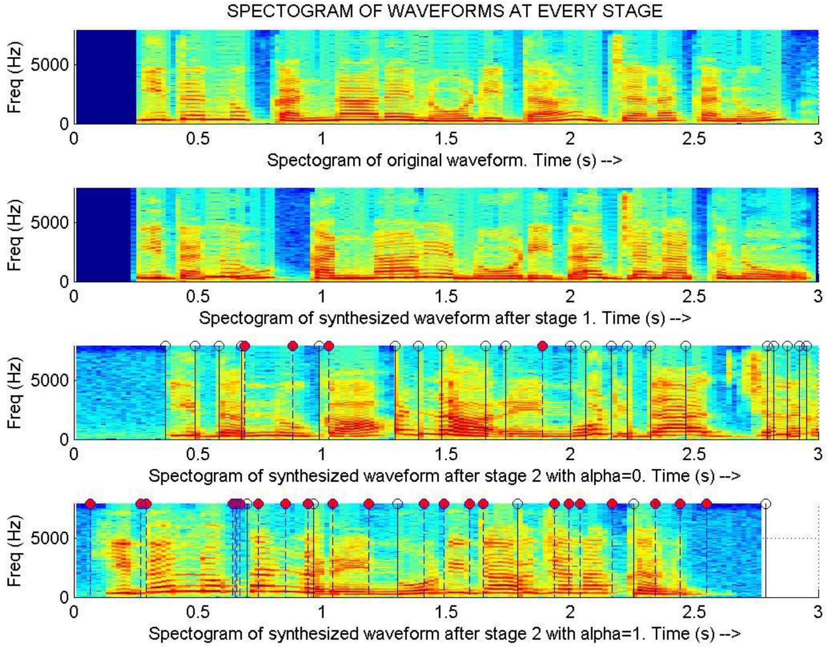


Fig. 6. *Spectrograms of a) reference speech, b) synthesized by conventional 1st-stage unit-selection, c) synthesized by 2nd-stage unit-selection for* $\alpha = 0$ *and d)* $\alpha = 1$

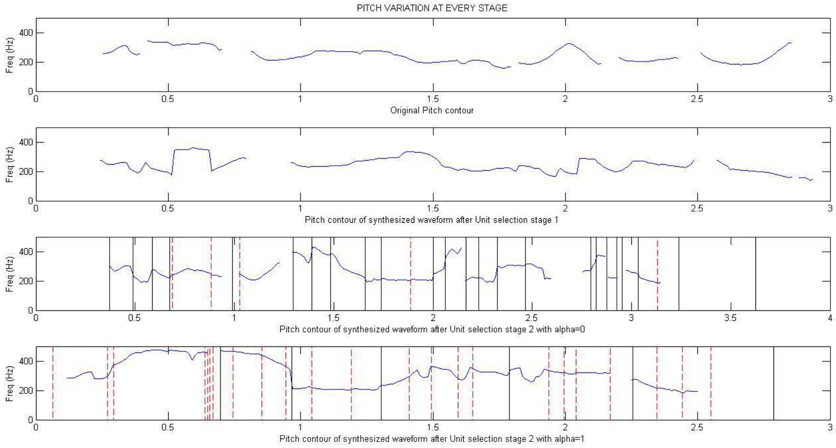


Fig. 7. *Pitch contours of a) reference speech, b) speech synthesized by conventional 1st-stage unit-selection, c) speech synthesized by 2nd-stage unit-selection for* $\alpha = 0$ *and d)* $\alpha = 1$

Note that the emphasis of our experiment is to see what effect the above behavior has on the prosodic transplantation. Hence, listening to the above 4 speech signals, revealed the following: i) the 2nd from top panel speech was a good synthesis - retaining neutral prosody from the 1st-stage unit-database; ii) the 3rd panel from top had poor perceptual quality, with highly audible spectral discontinuities and unit-overlaps, but with a semblance of prosody from the 2nd stage unit-database, iii) the bottom panel was highly inteligible, with very natural sounding speech, and more importantly, with a prosody that was distinctly in the prosodic style of the 2nd stage unit-database, and in that speaker's voice. In order to show this visually to some extent, we plot the pitch contours for the same 4 cases (as in Fig. 6) in Fig. 7. It can be noted that the pitch contours of top and next to top panels are those of the reference speech and as is found in the 1st stage unit-database. Of importance to our discussion are the bottom two panels, which show an intonation (pitch) contour markedly distinct from those of the top 2 panels, showing how different the speech synthesized by 2nd stage speech is from the 1st stage speech, and importantly, the bottom two panels show that the intonation contour in the 3rd panel from top is highly fragmented (causing annoying and unnatural perceptual pitch discontinuities) and a very perceptually pleasing intonation contour in the last panel at the bottom, marked by long continuous pitch contours corresponding to the long contiguous groups of units selected by the 2nd stage unit-selection with $\alpha = 1$.

This transfer of intonation, intensity and duration of the units from the 2nd stage unit-database carries the prosody inherent in the 2nd stage prosody database and clearly demonstrates the efficacy of the proposed framework to perform prosody transplanatation, further confirming the early findings of [4]. We expect further optimizations from the experimental scenarios outlined in the preceding section can lead to effective prosody control and incorporation strategies.

V. Conclusion

We have proposed a 2-stage unit-selection framework for prosody transplantation in TTS. This framework maps the unit sequence determined in a 1st-stage conventional unit-selection to an optimal sequence of units in a 2nd-stage prosodic unit-selection. We showed that the retrieved units from the 2nd-stage unit-database has an associated prosody in the prosodic-style of the 2nd-stage unit-database which can be further incorporated on to the spectral information of the 1st-stage unit-selection in any of the conventional synthesizer system. We have introduced a Viterbi algorithm for optimal unit-selection on pre-grouped units and have identified several experimental scenarios that will impact the nature of the prosody that can be transplanted from the 2nd-stage unit-selection. We have shown preliminary results that confirm the validity of the framework, and the promise of such a prosody transplantation scheme.

VI. Acknowledgment

The authors would like to thank the Department of Information Technology, Ministry of Communication and Technology, Government of India, for funding the TTS Consortium project, “Development of Text-to-Speech synthesis for Indian Languages Phase II”, Ref. no. 11(7)/2011HCC(TDIL). The work reported in this paper is part of the focus-topic at PESIT-BSC for 2015 under this funding.

References

[1] M. Bulut, S. Narayanan and L. Johnson. Synthesizing expressive speech overview: challenges, and open questions. Ch. 9, in Text to Speech Synthesis: New Paradigms and Advances, ed. By S. Narayanan, A. Alwan (Pearson Education, Prentice Hall, pp. 197-223, 2005.

[2] E. Eide, R. Bakis, W. Hamza and J. F. Pitrelli. Toward expressive synthetic speech. Ch. 11, in Text to Speech Synthesis: New Paradigms and Advances, ed. By S. Narayanan, A. Alwan (Pearson Education, Prentice Hall, pp. 241-272, 2005.

[3] K. Sreenivasa Rao Predicting Prosody from Text for Text-to-Speech Synthesis. Springer Brief, 2012.

[4] R. Prudon, Christophe DAlessandro and P. B. de Mareuil. Unit selection synthesis of prosody: evaluation using diphone transplantation. Ch. 10, in Text to Speech Synthesis: New Paradigms and Advances, ed. By S. Narayanan, A. Alwan (Pearson Education, Prentice Hall, 2005), pp. 225-239.