2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Phonetically conditioned prosody transplantation for TTS: Unit granularity, context and prosody styles

M. G. Khanum Noor Fathima, Mythri Thippareddy, M. Arunakumari, H. C. Mamatha, H. N. Supriya, A. Sricharan, V. Ramasubramanian*
PES Institute of Technology - Bangalore South Campus (PESIT-BSC)
Bangalore, India
*v.ramasubramanian@pes.edu

Abstract — **Recently, a 'phonetically conditioned' prosody-selection was proposed to realize a target prosody in a conventional TTS output unit-sequence. Here, target prosody-profiles were identified conditioned on their underlying phonetic content with varying degrees of matching and over variable length time-scales that are long as possible. In this paper, we extend these results in two ways: a) we consider various unit-definitions in the 2nd-stage unit-selection with respect to their relative influence on the phonetic conditioning, and hence the nature of prosody selected and, b) show the applicability of the proposed framework to a class of prosodic-styles. The various unit definitions include different granularity (such as fine-category phones to various broad-category phones) and various contextual effects, varying from no context to immediate left-/right-contexts (triphones) and quad-phone contexts. The performance of the proposed framework is characterized using various objective measures, and bring out the effect of the unit-definitions on the prosody selection, and the viability of the framework for a range of prosody styles.**

Keywords— Prosody selection, transplantation, phonetic conditioning, unit granularity, unit context, prosody styles

I. INTRODUCTION

Prosody in speech plays a range of functions, such as sentence mode, narrow focus word-stress by pitch-accent or dynamic-accent, prosodic styles (e.g. story-telling, news-reading) and emotional speech. Among such perceptual effect of prosody, the notion of prosody style is somewhat poorly defined, since it is a highly supra-segmental phenomenon, possibly at even longer time scales, marked by a categorical quality that this speech has in being considered a specific style, even without necessarily attending to shorter-time scale prosodic effects such as word-stress or sentence mode (e.g. interrogative sentence with a rising pitch in the end of the sentence). This perceptual cuing of a particular prosody style is therefore pervasively and consistently present through the entire speech corpus considered to be in a particular style, and it is difficult to characterize a set of rules that can produce a desired style in synthesized speech, such as is possible for word-stress.

Keeping this in view, in this paper we examine a framework for synthesizing speech with a desired 'prosody style', using the simple concept that unit-selection based concatenative synthesis from a unit-database realizes a so-called 'as is' prosody, i.e., it preserves and transfers the prosody style of the underlying unit-database on to the synthesized speech, a phenomenon now well acknowledged [1].

We first provide the main basis of this framework, and further examine a few scenarios by which such a framework can be realized, out of which we identify the most promising scenario, which we had proposed and studied earlier [2,3]. Within this framework, this paper examines issues related to unit-definition and the generalizability of the framework to several different prosody styles.

A. Basis of proposed framework

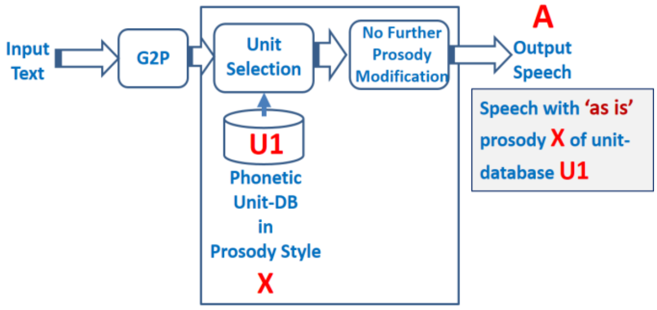
The framework proposed by us earlier (Mythri et al. [2,3]), termed 'phonetically conditioned 2-stage prosody transplantation' is along the lines of what was identified by van Santen et al., as 'Future approaches' in the review article [1] (Sec. 23.5, specifically 'hybrid approaches' in Sec. 23.5.1) from which we quote verbatim the following passage: "In order for unit selection synthesis to generate speech with natural and acceptable prosody, a different approach is needed. Some research has been devoted to select natural prosodic contours from dedicated prosodically balanced speech corpora in conjunction with the traditional unit search in a phonetically balanced speech corpus [4], [5], [6]. This significantly cuts down on the amount of speech data needed, as there is more emphasis on the phonetic context for the unit selection and more emphasis on the prosodic context for the prosody selection. The natural prosody contours are imposed on the units, so a high-quality speech modification algorithm is key for natural sounding speech. The use of dedicated prosody corpora allows for including different speaking styles and affective modes for synthesis". We note however, that despite such a pointed highlighting of this framework in the above review article, this suggested approach has not received any attention so far in the literature, and we could be among the first to address this in some detail.

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

To provide a more elaborate reasoning behind how the principle enunciated above can be realized in the form of a 2-stage framework, as we proposed and studied earlier [2,3], and which forms the basis for further enquiry in this paper, we take up in the following three different scenarios (or configuration) in which this principle can be put to use.

1) 'As is' prosody

Fig. 1 shows the structure of a conventional unit-selection based concatenative synthesis using a unit-database U1 which has a coherent prosody-style (e.g. news-reading style, i.e., speech of a news-caster reading news, and that is acknowledged to be in a style characteristic of news-reading). The speech output of this system also has the prosody style X of the database U1, in what is referred to as 'as is' prosody [1]. Now, if it is desired to change the prosody style of the output speech to style Y (e,g, story-telling), it becomes necessary to use a unit database, say U2, which is ideally spoken by the same speaker (for consistency in switching between style X and Y seamlessly). More importantly, since U1 played the role of a unit-database that provided both 'phonetic' and 'prosodic' coverage, likewise U2 also needs to be 'large' to provide 'phonetic coverage' even while being in style Y. Thus, in general, if it is desired to have speech output in arbitrary styles, it becomes necessary to create and annotate 'large' unit-databases in each of the target styles, even while providing phonetic coverage for good quality speech in the desired target unit-selection which uses 'units' derived and defined from the 1st-stage output as input and on which is performed a unit-selection from a 2nd-stage unit-database U2 (which is in the target prosody style Y) and yielding speech units (marked "B") which are 'as is' prosody of U2's style Y.

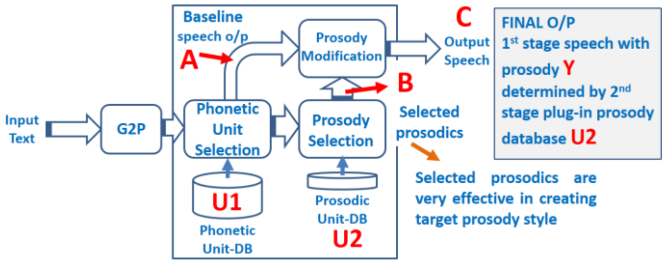


Speech with 'as is' prosody X of unit-database U1

Figure 1. Conventional unit-selection with 'as is' prosody style.

2) 2-stage unit-selection for prosody-selection and transplantation (Type I)

Fig. 2 shows a 2-stage framework, wherein the 1st stage is the conventional unit-selection based concatenative synthesis using a unit-database U1 in some prosody style. Let us assume this is in some 'neutral' style, i.e., speech spoken in declarative mode, for example. Thus, the output of this 1st stage will be of high quality, but with the 'as is' prosody of the unit-database U1 (e.g. neutral style, as assumed above). However, it is of interest to realize this speech in different styles, on demand (e.g. as dictated by prediction from text or reading modes, as in a single system which can be deployed in a variety of synthesis applications, such as for creating audio-books, or reading news, or creating emotional animation voice-overs, or interactive spoken dialog with conversational style, or classroom lecture styles etc.). For this, our earlier work [2,3], proposed a 2nd-stage



FINAL O/P 1st stage speech with prosody Y determined by 2nd stage plug-in prosody database U2

Selected prosodics are very effective in creating target prosody style

Figure 2. 2-stage unit-selection Type I

Some of the important aspects of this scenario are as follows:

1. The 1st stage provides the speech units (i.e., speech waveforms or short-time spectra) from U1 in a way faithful to the phonetic units derived from the G2P of input text.
2. The 2nd stage provides the prosody associated with the units selected by the 2nd stage unit-selection from U2. This is essentially a 'prosody selection', conditioned on the units of 1st stage output (i.e., the input text). The selected prosody is typically in the form of duration, F0 contours and gain profiles (marked "B") of the individual units. This prosody is further transplanted on to the speech units of 1st stage, for instance, by prosody modification of the speech units of 1st stage according to the prosody selected by the 2nd stage employing TD-PSOLA, MBROLA or STRAIGHT kind of techniques to yield the final speech output, marked "C", which is in the voice of the speaker of U1, but with the target prosody Y of U2, i.e., the prosody Y has overridden the (neutral) prosody X of A.
3. As outlined in Sec. 1.A, U2 need not have a phonetic coverage since its objective is to 'select' units that match with the unit-sequence at the output of 1st stage, with the sole purpose of deriving the 'as is' prosody of U2 that is 'associated' with the selected units. By this, U2 needs to have prosodic-coverage, in terms of its units, in terms of adequate contextual occurrences, which have the characteristic prosody style consistently.
4. By the above, the 2nd-stage unit-database U2 need not be as large as the 1st-stage unit-database U1, and can be smaller, in line with the observation in Sec. 1.A '…this significantly cuts down on the amount of speech data needed…', implying a relatively *small* size of U2.
5. This leads to the desired property of the 2-stage framework that there need be only one large U1, providing adequate phonetic coverage for a high-quality speech marked "A", and U2 can be a *small* "plug-in" database, one per desired target prosody style, possibly by the same speaker as U1 or other speakers. This kind of 'plug-in' prosody results in a system that allows easy deployment of such a system to be tailored (or even switched seamlessly) to several desired prosody styles, by merely requiring a small U2 (recorded and annotated a priori) to be plugged into the 2nd stage.

3) 2-stage unit-selection for prosody-selection and transplantation (Type II)

Note that the above description of the Type I form of the 2-stage framework naturally leads to asking whether the kind of configuration in Fig. 3 might also be more appropriate. Here, the 1st and 2nd stage unit-selections operate in 'parallel', i.e., the G2P output (e.g. phonetic unit sequence) can be fed in 'parallel' to the two stages, the 1st stage yielding the speech units (e.g. waveforms) from U1 and the 2nd stage yields the desired 'as is' prosody Y from U2, which are then combined by a prosody modification module, i.e., modifying the waveforms or spectra of "A" by the selected prosody "B" to yield the final speech output "C" in the voice of the speaker of U1 but with the prosody of U2.

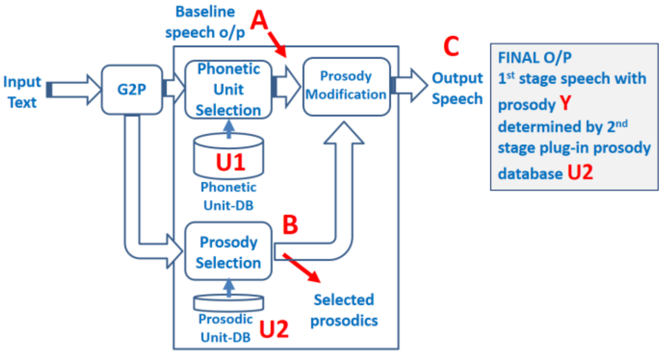


Figure 3. 2-stage unit-selection Type II

However, while operationally the two types Type I and Type II appear same, Type I has a crucial advantage over Type II, in the sense that the units fed to the 2nd stage unit-selection in Type I can have 'linguistic features' that are faithful to and derived from the input text and also 'acoustic features' that can be derived from the 1st stage output (i.e., the speech signal) by signal processing techniques (e.g., spectral features, spectral dynamics, prosodic features such as duration, intensity and F0 etc.), so that the 2nd-stage unit-selection can perform a 'prosody selection' conditioned on a rich set of features, that can control the degree of how close the selected prosody is to the desired perceptual prosody from U2.

The earlier work of Prudon et al. [7] falls within Type-II kind of framework, where the prosodic parameters ("B" in Fig. 3) are 'selected' directly from input text (after a G2P). The prosodic parameters are primarily the duration of units, intonation and gain contours. These were then combined with the diphone units (acoustic information, at "A" in Fig. 3) derived from the text (after G2P) and synthesized as "C" using MBROLA. This allowed circumvention of syntactic analysis and syntactic-prosody rules, as the system used the 'prosodic part' of a specified prosody corpus directly.

From the above reasoning, we choose to use Type I in the work here. To summarize from the above discussion, the framework proposed by us earlier [4] and which is examined here in more detail is the configuration '2-stage unit-selection' Type I, which performs a 'phonetic conditioning', i.e., which matches the units output by the 1st-stage to the units selected from the prosodic-database U2 to varying degrees, depending on i) the nature of units so defined, and ii) the efficacy of the 2nd-stage unit-selection, particularly with respect to the definition of the unit-cost (relative weights across the unit vector components) and the relative weighing between unit-cost and concatenation cost. These two in turn directly controls the 'degree of unit match' by the 2nd stage unit-selection – in terms of to what degree the phonetic match is realized, and how long the contiguity of the selected units are, and thereby the associated prosodic-profiles – with the expected result that higher the degree of unit-match which are as long as possible, the prosody associated with the resultant units from the prosodic-database will be the desired prosody in the sense that 'it is the same text (i.e., phonetic sequence) as the input, but spoken with the prosody-style Y of the target 'prosodic-database' U2'.

For realizing a prosody 'style', we show such a 'phonetic conditioning' to be a kind of 'sufficient, but not necessary' condition, where long-contiguity, fine-grained unit conditioning is 'sufficient' for a good prosody-selection, while being 'not necessary', since there can be other means of 'selecting' the signature prosodic features (of a desired target prosody) that can be transfer to the 'carrier' speech "A" in Fig. 2, to yield an equivalent perception of the desired prosody style in the final output "C".

We explore in this paper the following:

a) various unit-definitions in the 2nd-stage unit-selection to establish their influence on the phonetic conditioning, and hence the nature of prosody transplanted and,
b) the applicability of the proposed framework to a class of five prosodic-styles, namely, neutral speech, story-telling, drama, political discourse and classroom lecture.

II. 1ST STAGE AND 2ND STAGE UNIT-SELECTION

Fig. 4 shows the detailed schematic of the framework (Type I in Fig. 2) whose basis was referred above and which is examined in further detail in this paper.

The 1st-stage unit-selection is the conventional part of a concatenative TTS, e.g. the grouped monophone-only Viterbi, as part of a mixed-Viterbi formalism proposed by us recently [8]. We describe here the 2nd-stage prosody transplantation in Fig. 4 in some detail.

Let $O = (\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_t, \dots, \boldsymbol{o}_T)$ be the phonetic unit-sequence at the output of the 1st-stage unit-selection, where $\boldsymbol{o}_t$ is a vector with components made of the primary phonetic unit label and any other contextual features of the unit selected from the 1st-stage unit-database, as relevant for the 2nd-stage prosody-selection (e.g. linguistic features as in [9]). Note that $\boldsymbol{o}_t$ can afford to have features derived by signal-processing on the 1st-stage and 2nd-stage unit-databases i.e., not necessarily from only text.

Let $\mathcal{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_i, \dots, \boldsymbol{u}_N)$ be the 2nd-stage unit-database, made of a sequence of N phonetic units, with durations of the associated acoustic segments as $l_1, l_2, \dots, l_i, \dots, l_N$. For a given input unit-sequence $O = (\boldsymbol{o}_1, \boldsymbol{o}_2, \dots, \boldsymbol{o}_t, \dots, \boldsymbol{o}_T)$ the 2nd-stage unit-selection finds the optimal unit sequence $U^* = (\boldsymbol{u}_{q_1^*}, \boldsymbol{u}_{q_2^*}, \dots, \boldsymbol{u}_{q_t^*}, \dots, \boldsymbol{u}_{q_T^*})$ where $\boldsymbol{u}_{q_t^*} \in \mathcal{U}$ by determining the optimal unit indices $Q^* = (q_1^*, q_2^*, \dots, q_t^*, \dots, q_T^*)$ that minimizes the distortion between O and any U (from the 2nd-stage unit-database) as given by,

$$Q^* = arg \min_{Q}(1-\alpha)\sum_{t=1}^{T} D_u\left(\boldsymbol{o}_t, \boldsymbol{u}_{q_t}\right) + \alpha\sum_{t=2}^{T} D_c(q_{t-1}, q_t) \quad (1)$$

Here, $D_c(q_{t-1}, q_t)$ is the concatenation cost that is 0 if $\boldsymbol{u}_{q_{t-1}}$ and $\boldsymbol{u}_{q_t}$are contiguous in the 2nd-stage unit database, and is the Euclidean distortion between the last frame of the acoustic segment $\boldsymbol{u}_{q_{t-1}}$and the first frame of the acoustic segment $\boldsymbol{u}_{q_t}$, if they are not contiguous.

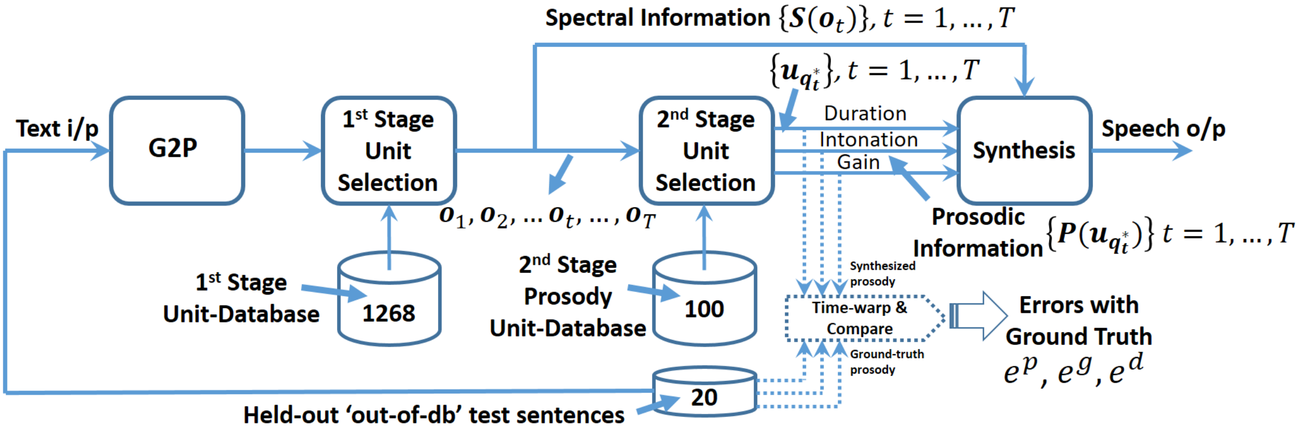

Spectral Information $\{\boldsymbol{S}(\boldsymbol{o}_t)\}, t = 1, \dots, T$
$\{\boldsymbol{u}_{q_t^*}\}, t = 1, \dots, T$
Text i/p
G2P
1st Stage Unit Selection
2nd Stage Unit Selection
Duration
Intonation
Gain
Synthesis
Speech o/p
$\boldsymbol{o}_1, \boldsymbol{o}_2, \dots \boldsymbol{o}_t, \dots, \boldsymbol{o}_T$
Prosodic Information $\{\boldsymbol{P}(\boldsymbol{u}_{q_t^*})\}\ t = 1, \dots, T$
1st Stage Unit-Database
1268
2nd Stage Prosody Unit-Database
100
Synthesized prosody
Time-warp & Compare
Ground-truth prosody
Errors with Ground Truth
e^p, e^g, e^d
20
Held-out 'out-of-db' test sentences

Figure 4. Schematic of the 2-stage unit-selection framework proposed and studied here

The above solution for the optimal unit indices Q^* is solved using Viterbi unit-selection typically using a decision tree. In the present work, we have used a more unconstrained version of the decision tree, which is equivalent to using only the root of the decision tree corresponding to groups of units G_t which have all the units in the database that have the same phonetic identity as the unit $\boldsymbol{o}_t$, in the input unit sequence. This yields the best quality of unit-selection, being a super-set of any leaf in a conventional decision tree, though at the expense of computation time, but yielding the best phonetic-conditioning from an unconstrained unit-selection.

This yields the optimal unit sequence U^* from the 2nd-stage corresponding to the input unit-sequence O 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 $\boldsymbol{o}_t$ and $\boldsymbol{u}_{q_t^*}$ (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, it can be expected that the prosody associated with U^* will be typical of the style of the 2nd-stage unit-database, under the condition that the underlying phonetic units match exactly (which is ensured by the grouping mentioned above) and more importantly, by the degree of contiguity which a choice of α allows (i.e., with $0 \leq \alpha \leq 1$, larger values of α yield an unit sequence U^* with longer contiguity, in the process, retrieving a long naturally articulated prosody associated with such long sequences of units)

Once U^* is determined above for a given input O, the prosodic transplantation is done as follows (and as is shown in Fig. 4): Let $S(\boldsymbol{o}_t)$ be the speech waveform of unit $\boldsymbol{o}_t$ and $P\left(\boldsymbol{u}_{q_t^*}\right)$ be the set of prosodic parameters (duration of unit $\boldsymbol{u}_{q_t^*}$ and its gain and pitch contours). These can be used in TD-PSOLA or STRAIGHT kind of framework to perform time-scale, pitch-scale and short-term gain modifications to yield the desired prosody transplantation. In this work, we have used STRAIGHT based prosody modification with individual and independent control of the individual 'prosodics', i.e. F_0 contour, gain contour and unit-durations.

III. Unit Definition

We consider here the following four scenarios in the definition of units that can be expected to influence the degree of match between $\boldsymbol{o}_t$ and any unit $\boldsymbol{u}_{q_t} \in G_t$ (i.e., the degree of 'phonetic-conditioning':

1. The granularity of the units can span fine- to course-grained representations, such as, i) acoustic representation (e.g. MFCC) at one extreme (as can be derived from the signal, i.e., from the 1st stage output and the 2nd stage unit-database) possibly representing the strongest phonetic-conditioning, ii) fine-category phoneme class (as in the results in the next section) or, iii) coarse-grained, such as various levels of broad-category phoneme classes,
2. The contextual span of the unit, in any of the granularity above, such as various extents of left- and right-context information. This reflects in the choice of units in G_t and the unit cost definition and controls the degree of conditioning through context-dependency,

3. Prosodic features, such as duration, intensity and F0 can also be included into the unit vector definition, with corresponding weighting in the unit-cost. This can provide selective control on their relative influence on the resultant prosodic-copy and,
4. The notion of 'phonetic'-conditioning can be generalized to other types of conditioning, such as on linguistic-features, (as can be derived from text, as is done for text-to-prosody mapping [9]), structural features [10] or, acoustic-features (as can be derived from the signal itself). This can help going beyond only prosodic 'style' copy, towards being able to realize other prosodic-effects such as narrow-focus word-stress or, in general, expressive speech closely governed by the text content.

Of these, we focus on definitions 1 and 2 here (specifically, 1(ii) and 1(iii)), to explore the different definitions of the unit underlying the 'phonetic' (i.e. 'unit') conditioning within the framework outlined above.

IV. Experiments and Results

We present results on an Indian language 'Kannada' (one of the 22 scheduled languages of India). The 1st-stage unit-selection uses a neutral prosody-style database of 1268 sentences (~4 hrs). The 5 prosodic styles we have considered are *neutral speech, story-telling, drama, political discourse and classroom lecture.* For each of these 'styles', the 2nd-stage prosodic-database corresponds to 100 sentences of that style, (~15 min); (note that we had shown earlier [3] that it is possible realize effective prosody selection oration even when the 2nd-stage prosody database is reduced from 45 min to 15 min).

As shown in Fig. 4, 100 sentences of each of the 5 prosody styles are used as the 2nd-stage unit-database, and an additional 20 sentences of each style are used as 'out-of-database' (out-of-db) sentences; these 20 sentences provide the ground-truth prosody against which to compare the synthetic prosodic contours, i.e., F0 or pitch (p), gain (g) and duration (d) profiles, to yield the respective errors e^p, e^g and e^d – which are computed between time-normalized synthesized and ground-truth prosody contours. 20 sentences from within the 100-sentence prosody-database are used as 'in-database' (in-db) sentences, to be able to calibrate the system, and provide the baseline performance for the out-of-db sentences for any of the experiments to follow.

A. *Fine- to course-grained units*

In line with the two unit-definitions identified above for the present work, we examine the nature of units of the following types: i) random choice (labeled 'rnd'), where an index $\boldsymbol{u}_{q_t^*}$ is selected randomly from the group G_t (under the unit-definition fc-lc-rc as defined in Item (ii) next) to establish a base-line (worst-case) performance when the phonetic conditioning is arbitrary without any care for preserving contiguity and hence the naturalness of the associated prosodic parameters manifesting as large errors e^p, e^g and e^d; this clearly validates the need and viability of phonetic conditioning to bring about a good prosody transplantation, as discussed in Sec. I, ii) fine-category with left- and right-context included in the unit vector and associated unit-cost (fc-lc-rc), iii) fine-category with no left- and right-context (fc-no-lc-rc), iv) fine-category with quad-context (fc-llc-rrc), v) broad-categories varying from vowel/consonant coarse-category one end (bc1), to two other broad category groupings (bc2 and bc3) with bc3 having more units than bc2 and hence being closer to the fine category.

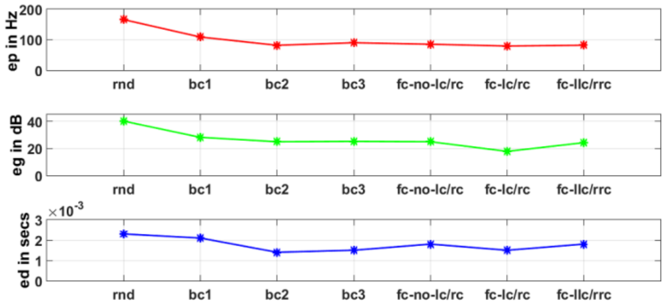


Figure 5. Pitch (F0), gain and duration errors - e^p, e^g and e^d – for various unit granularity

Fig. 5 shows $e^p, e^g$and e^d for these unit definitions. As could be expected, the 'rnd' case performs the worst, with considerably large errors $e^p, e^g$and e^d than the other unit definitions – thereby underscoring the importance of phonetic conditioning to provide an appropriate prosody transplantation. Coarse category units perform poorer than fine-category given that the degree of phonetic conditioning is more relaxed the more broader the categorization, but nevertheless, it is worth noting that the increase in the errors are marginal and gives the interesting possibility that one could even work with coarse-grained units, and still get a fair degree of prosody transplantation, with associated lower complexity in the 2nd-stage decoding, as well as (and more importantly), in the significantly lower effort needed to perform the segmentation and labeling of the unit database into coarser units for a given accuracy. Increase in the context-span helps reduce the error, keeping with the notion that longer-span contexts represent a tighter phonetic-conditioning, which in turn result in more contiguous choice of units and hence more naturally co-articulated prosody to be transferred.

B. *Effect of α*

The weighting factor α (trading off unit cost against concatenation cost), controls the degree of match between the

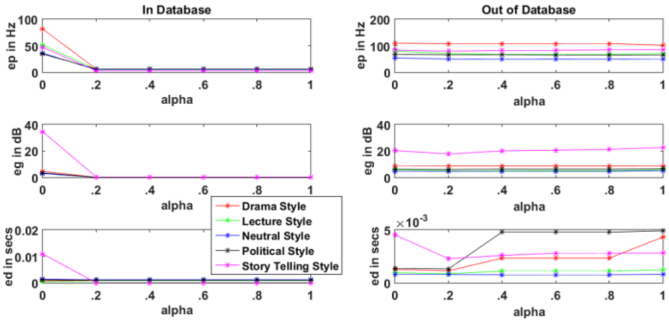


Figure 6. Pitch (F0), gain and duration errors – e^p, e^g and e^d – as a function of α for the 5 prosody styles for in-db sentences (left-panel) and out-of-db sentences (right-panel)

primary phonetic identity between the units $\boldsymbol{o}_t$ and $\boldsymbol{u}_{q_t^*}$. 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. To this end, Fig. 6 shows the effect of α for in-db and out-of-db cases on $e^p, e^g$and e^d for all the 5 prosody styles (averaged over the 20 sentences in each case). The in-db errors are close to 0 for > 0 , as it should be, since the unit-selection selects a single long sequence of the in-db units. For $\alpha = 0$, the contiguous units are not selected, since the concatenation cost (weighted by α) is 0, and the decoding selects units arbitrarily from the respective unit-groups, leading to fragmentation and mismatch of the retrieved acoustic segments. This causes the higher errors for all the three errors at $\alpha = 0$. The out-of-db errors are higher than the in-db errors understandably, and are of comparable values for all the 5 prosody styles, indicating that each of the styles has transferred effectively even for out-of-db sentences, consistent with the results we reported earlier [3] for one prosody style. This generalizability of the proposed framework for a class of distinct prosody styles is a positive outcome.

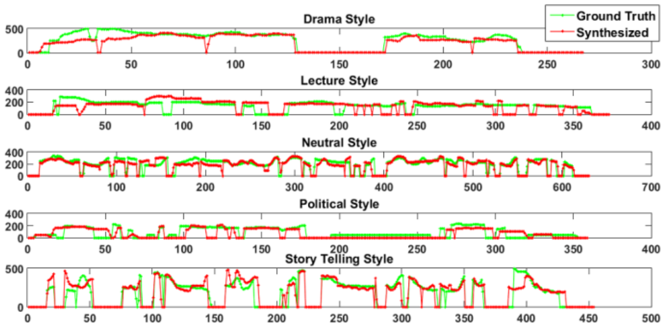


Figure 7. Pitch contours - synthesized and ground-truth - for one out-of-db sentence for all 5 prosody styles

C. F0-contour comparison

Fig. 7 is a panel of 5 plots for one out-of-db sentence, showing the ground-truth and synthesized (and time-warped) F0 contours, for the 5 prosody styles. The F0 contours of synthesized speech for all the 5 styles bear a remarkably close match to the ground truth F0 contours, validating the premise that when the degree of phonetic-sequence match is good, the corresponding associated prosody turns out to be a good copy of the prosodic-style in which these sequence of units 'would' have been spoken in the target prosody style.

V. Conclusions

We have examined a 2-stage unit-selection based prosody selection and transplantation framework for realizing a desired target prosody made available in the form of a plug-in 'prosody database'. The 2nd-stage unit-selection performs a 'phonetic conditioning', realized in terms of match between the phones in the 1st stage output sequence and the units in the prosody database. We have demonstrated the need for and importance of such a phonetic conditioning by showing poor performance for no conditioning, (e.g. a random conditioning) and have focused on various definitions of the units for such a phonetic-conditioning, such as the granularity of units and different contextual spans. We have also shown that the method is generic enough to generalize to a class of distinct prosody styles.

Acknowledgment

The authors 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).

References

[1] J. van Santen, T. Mishra and E. Klabbers, "Prosodic processing", Ch. 23, pp. 471 - 487, in Springer Handbook of Speech Processing, eds. Benesty, Sondhi and Huang, 2008.

[2] Mythri Thippareddy and V. Ramasubramanian, "Prosody transplantation using unit-selection: Principles and early results", IEEE CONECCT '15, Bangalore, Jul 2015.

[3] Mythri Thippareddy, Noor Fathima, D. N. Krishna, Sricharan, V. Ramasubramanian, "Phonetically conditioned prosody transplantation for TTS: 2-stage phone-level unit-selection framework", In Proc. Speech Prosody 2016, pp. 781-785, May-June 2016, Boston, MA.

[4] F. Campillo-Daz and E. R. Banga, "Combined prosody and candidate unit selections for corpus-based text-to-speech systems", Proc. of the 7th International Conference on Spoken Language Processing, Denver, CO, pp. 141-144, 2002.

[5] A. Raux and A. Black, "A Unit Selection Approach to F0 Modeling and Its Application to Emphasis", In ASRU 2003, St. Thomas, US Virgin Is., 2003.

[6] J. van Santen, A. Kain, E. Klabbers and T. Mishra, "Synthesis of prosody using multi-level sequence units", In Speech Communication, 46(3-4), pp. 365-375, 2005.

[7] R. Prudon, Christophe D'Alessandro 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.

[8] Mythri Thippareddy, C. Mahima, S. Adithya, Sunil Rao and V. Ramasubramanian, "G2P-free grapheme-to-speech synthesis: UTF-8 based automatic unit-database annotation and mixed Viterbi unit-selection", Proc. O-COCOSDA '15, Shanghai, China, Oct. 2015.

[9] H. Mixdorff, "An integrated approach to modeling German prosody", Post-doc Thesis, Universitats verlag, Dresden, Germany, 2002.

[10] K. Raghava Krishnan, "Prosodic Analysis of Indian Languages and its Applications to Text to Speech Synthesis", M. S. Thesis, Dept. Electrical Engineering, IIT-Madras, Chennai, India.