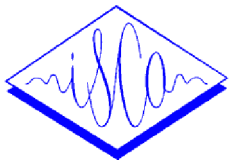
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Usages of an external duration model for HMM-based speech synthesis.

Javier Latorre[1,2], *Sabine Buchholz*[1], *Masami Akamine*[2]

[1]Toshiba Research Europe, UK
[2]Toshiba Corporate Research & Development Center, Japan
javier.latorre@crl.toshiba.co.uk, sabine.buchholz@crl.toshiba.co.uk, masa.akamine@toshiba.co.jp

Abstract

In this paper we analyze three different approaches to improving the quality of an HMM-based speech synthesizer by means of an external duration model. The first approach uses the external duration model in a standard way to define the phone duration during synthesis. The second is a novel approach that uses the phone duration to create additional context features for the decision trees clustering. The third is a combination of the previous two approaches. A subjective evaluation showed a quality improvement with respect to the baseline for all three approaches, although for differing reasons. The standard approach produces an improvement in the duration estimation. The second approach degrades the duration estimation but improves the logF0 and aperiodicity by better modeling of their dependencies with respect to the duration. Finally, the combined approach benefits from the improvements of the other two and yields the best result of ca. 16% higher preference than the baseline among native English speakers.

Index Terms: speech synthesis, prosody, duration, HMM-based, external duration model

1. Introduction

Duration, pitch and power are the three main components of the prosodic signal. The duration defines the rhythm and speed in which a sentence is spoken: which sounds are longer, which shorter, which part of the sentence has to be pronounced faster, which slower, etc. The most important factors for the duration are phonological: phonetic sequence, stress, etc. However, speakers also use the duration to communicate other types of information such as the structure of the sentence, which part of the discourse is most important, etc. As such, other factors such as syntax and semantics need to be considered as well in order to correctly estimate the phone duration of an utterance. For a text-to-speech (TTS) system, this implies that an accurate duration model is essential in order to synthesize natural and intelligible speech. Many types of machine learning techniques have been applied to predict the speech durations: Bayesian networks [1], linear statistical models [2],[3], neural networks [4], classification and regression trees [5], etc. In this paper we study duration modeling in the framework of HMM-based speech synthesis [6], and the possibility to improve it by using an external duration model.

The rest of the paper is organized as follows. Section 2 describes how the duration is modelled in HMM-based speech synthesis, its problems, and the standard way to integrate an external duration model. Section 3 explains a new approach to integrate an external duration model into HMM. Section 4 describes a subjective evaluation we conducted to assess the effectiveness of the different approaches. Section 5 discusses the possible causes of the results and finally in section 6 we draw conclusions based upon these results.

2. HMM duration model

In HMM-based synthesis, F0, duration, spectrum and aperiodicity are modelled simultaneously at the frame level. In its original implementation [6], the duration model was created at the end of the training process, by clustering in a decision tree the occupancy statistics of the acoustic models obtained in the last expectation-maximization (EM) iteration. This way of modeling the duration introduced a lack of consistency between training and synthesis: the state duration was modelled by an exponential distribution during training but a Gaussian distribution during synthesis. This inconsistency was solved in HSMM [7] by integrating the Gaussian model of the state duration within the EM training. Although this modification helps to improves the total quality, the duration estimated by HSMM is still worse than the one estimated by other machine learning methods. In a preliminary experiment, we compared the duration predicted by a 3-state HSMM model and a a Quantification method type 1 (QMT1) model [8]. The root mean square error (RMSE) and the relative RMSE (percentage of the error with respect to the phone duration) that we obtained were 34.25ms and 31.15% for HSMM versus 27.5 ms and 29.5% for the QMT1 model.

If we consider the manner of modeling the duration in the HMM/HSMM framework we find two problems. The first is that although the duration is a prosodic signal, it is modelled at a sub-segmental level (the state). No constraints at a higher level are considered, and supra-segmental information is used only implicitly in the decision tree clustering. The second problem is that in order to model the duration at a state-level, it is represented as a vectorial variable with a dimension equal to the number of states of the hidden Markov model. The reason for using such state-level model is that in HMM/HSMM-based synthesis, the duration model defines the state sequence that will be used to to synthesize the other streams (spectrum, aperiodicity and F0). During the clustering of the duration models, the error that gets minimized is the average vectorial distance of the cluster models and its centroid. At a segmental level, the duration of some phones does indeed contain some vectorial component, for example in the initial and final part of a diphthong, or the closure/burst/aspiration of plosives. However, from a prosodic point of view, duration is primarily consider as a scalar variable. This means that the main goal of the training process should be to reduce the error between the predicted scalar value and the real one. Although scalar and vectorial distance are strongly correlated, they are not the same. Figure 1 shows the difference between clustering based on a scalar and a vectorial dimension in a hypothetical two-state HMM s_1, s_2. If the models are clustered based on their vectorial value, d_1 will be clustered

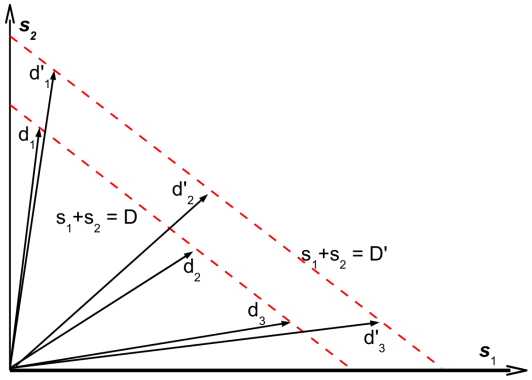


Figure 1: *Difference between clustering a vectorial and a scalar variable*

with d'_1, d_2 with d'_2 and d_3 with d'_3. However, if they are clustered based on their total duration $(s_1 + s_2)$ d_1,d_2 and d_3 will be in one cluster with total duration D and d'_1,d'_2 and d'_3 in the one with total duration D'. In order to alleviate this problem, other researchers have tried to combine the state-duration model with some external duration models at higher linguistic levels, such as the phone [9], or even longer units [10]. In these approaches, the models are combined by simultaneously maximizing the log-likelihoods of the HMM state-level model and the external higher-level model.

.

2.1. Explicit phone duration

To integrate an external duration model into HMM-synthesis, the standard way is to explicitly force the generation algorithm to use the phone duration predicted by that model [11]. In this way, the state-level duration model is only used to distribute the given phone duration among the states. The solution is obtained from the maximization of the log-likelihood of the state-duration vector $\boldsymbol{d}$ under the constraint

$$\sum_{s=1}^{S} \boldsymbol{d}_s = T \tag{1}$$

where S is the total number of states and T is the duration that we want for the S-states sequence, i.e. the external phone duration. Assuming that the state-duration model of a phone is Gaussian with mean $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$, the solution to this maximization is

$$\boldsymbol{d} = \boldsymbol{\mu} + \rho \cdot diag(\boldsymbol{\Sigma}) \tag{2}$$

with ρ

$$\rho = \frac{T - \sum_{s=1}^{S} \boldsymbol{\mu}_s}{trace(\boldsymbol{\Sigma})} \tag{3}$$

3. Implicit phone duration

In the HMM-based synthesis framework, the duration defines the state sequence but not the statistical model that should be associated with each state. In unit selection synthesis on the other hand, the duration is one of the most important factors in the selection of the units. The are two reasons for this. The first one is to reduce the distortion introduced by the signal processing required to modify the unit length. The second reason is because the characteristics of segmental and prosodic units are considered to vary depending on their length. The implicit-duration approach that we propose in this paper, is an attempt to emulate this into the HMM framework. The idea is to use the phone duration as an additional feature in the decision tree clustering of the duration, spectrum, aperiodicity and F0 models. In this way, we can model the dependencies between these streams and the total phone duration. In theory, a decision tree should be able to learn any dependency of the phone duration by using the same set of factors used to predict the duration. In practice, this only occurs when the amount of training data is infinite or very large. In the normal case the amount of data is limited. Therefore, when the decision tree arrives at a node where a duration question should be asked, it lacks the number of samples required for all the additional splits that would be needed to support the combination of factors on which the duration depends. In that sense, a duration context feature can be seen as a short-cut that encapsulates a complex combination of several other features.

The implementation of the implicit-duration approach in the current HMM framework is as follows: During the training, the phone durations are obtained from the force- or hand-alignments of the training data. These durations are then quantized and used to create features for the full-context labels that will be used in the decision tree clustering. In our implementation we quantize the phone duration in units of 10ms (2 frames).During synthesis, the duration values estimated by the external model are used to define the values of the duration features for the full-context labels. These labels are then used to select from the decision tree the sequence of acoustic models associated with the input sentence.

4. Experiment and results

4.1. Compared models

We conducted a subjective evaluation to test the effectiveness of three different ways to integrate an external duration model: explicit duration, implicit duration, and a combination of both. For the experiment we used HTS [12] to train two HSMM models: a baseline model based on the standard set of context features, and an implicit-duration model that adds to the full-context labels of the baseline phone duration features for current, previous and next phones. Both models were 5-states models. They were trained on a speech database of approximately 5.5 hours as spoken by a single US-English female speaker and sampled at 16kHz. The observation vector consisted of three streams: spectrum, aperiodicity and F0. The spectrum stream consisted of 80 coefficients, 40 LSP coefficients (including gain) and their delta. The LSP coefficients were calculated from a spectral envelope obtained in a pitch synchronous analysis and resampled with a 5ms. shift. The aperiodicity spectrum was obtained using the pitch-scaled harmonic filter algorithm (PSHF) [13] and encoded by a single parameter. This parameter indicates the frequency from which the spectrum can be considered fundamentally aperiodic. Its meaning is basically the same as the 'maximum voiced frequency' in the Harmonic plus Noise Model (HNM) [14]. The aperiodicity stream consisted of this parameter and its delta. Finally, the F0 stream consisted of the logarithm of F0 and its delta. The F0 stream was modelled using a multi-space distribution (MSD) [15]

4.2. Objective evaluation

In parallel to the subjective evaluation, we calculated the objective phone duration error over a held out subset of the data. For

Table 1: *Objective evaluation of the phone duration*

Model	Absolute RMSE	Relative RMSE
HSMM Baseline	26.9ms	28.7 %
QMT1	22.04ms	21.58%
Implicit:real input	11.42ms	10.45%
Implicit:QMT1 input	54.73ms	62.34%

the implicit-duration model we computed the errors with two types of duration input: real durations (non-realistic case), and the durations as predicted by the QMT1 model. Table 1 shows the results. The proportion between the RMSE of QMT1 and the HSMM baseline are similar to those mentioned in Section 2. The RMSE of the implicit-duration model with real duration corresponds to the quantization step. However, the RMSE of the proposed implicit-duration with QMT1 input is higher than we expected. The reason for this observation is the strong correlation between the scalar phone duration feature at the input and the vectorial output suggested by the low RMSE when using real duration. Due to such correlation, the RMSE of the implicit-duration model with QMT1 input becomes higher than the added RMSE of the QMT1 and HSMM baseline.

4.3. Subjective evaluation

4.3.1. Conditions

With each one of the two trained models, a set of 120 test sentences were synthesized in two ways, first with the phone duration provided by the HSMM state-level duration model, and second by explicitly forcing the phone duration to be the one estimated by the external QMT1 model. As a result we obtained four sets of stimuli. To evaluate the effectiveness of each approach we compared them in 4 subjective evaluations:

A **Baseline vs. Explicit**: Compare the speech generated by the baseline model with and without explicit phone duration.

B **Baseline vs. Implicit**: Compare the speech generated by the baseline and the implicit-duration model with the phone duration given by their respective state-level models.

C **Baseline vs. Implicit+Explicit**: Compare the speech generated by the baseline with its own phone duration, with the speech generated by the implicit-duration model with explicit phone duration.

D **Explicit vs. Implicit**: Compare the speech generated by the baseline model with explicit phone duration versus the implicit-duration model with its own duration.

For each evaluation a total of 10 subjects evaluated a set of 45 stimuli pairs in a preference test. They all were speech technology experts. Subjects were allowed to choose a "None" option if they judged both stimuli to be equal. For each subject the pair of stimuli were randomly selected from the collection of 120 test sentences. Six of the subjects of each evaluation were native English speakers, and the other 4 were highly proficient English speakers who have been living for several years in English speaking countries. The average sentence length was 8.15 words. The external duration model was an improved QMT1 model trained using feature selection [16]. We found that the phone durations predicted by the QMT1 model for the set of test sentences was on average 7.5% shorter than those predicted by the baseline duration model, and 3.74% shorter than those predicted by implicit-duration model.

4.3.2. Results

Tables 2 and 3 show the results of the subjective evaluations among all the subjects as well as among just the 6 native English speakers. On the table, **X** and **Y** refer to the first and second model mentioned in the evaluation conditions.

Table 2: *Results for all subjects*

Eval.	**X model**	**None**	**Y model**	**(X-Y) 95% margin**	**binary z-score err.**
A	21.1%	49.8%	29.1%	$8\% \pm 6.59\%$	4.48%
B	30.4%	30.7%	38.9%	$8.5\% \pm 7.73\%$	3.57%
C	28.9%	30.2%	40.9%	$12\% \pm 7.8\%$	0.55%
D	31.4%	32.1%	36.5%	$5.1\% \pm 8.04\%$	15.24%

Table 3: *Results for native English speaker*

Eval.	**X model**	**None**	**Y model**	**(X-Y) error 95% margin**	**binary z-score err.**
A	20.7%	46.7%	32.6%	$11.9\% \pm 8.82\%$	2.53%
B	33.3%	24.5%	42.4%	$8.9\% \pm 10.42\%$	7.18%
C	27.8%	28.5%	43.7%	$15.9\% \pm 10.26\%$	0.45%
D	32.4%	29.8%	37.8%	$5.4\% \pm 10.97\%$	20.9%

Though small, the differences between the baseline and the models that use external duration are significant in all cases except for the "Baseline vs. Implicit" case among native English speakers. Among native English speakers, the strongest improvement seems to come from the better duration estimation provided by the explicit usage of the QMT1 duration. This result is interesting if we consider that the RMSE of the QMT1 model is less than one frame-shift better than that of the baseline. Moreover, given this preference for such a duration difference, we should have had much lower preference for the implicit-duration model in "Baseline vs. Implicit" and "Explicit vs. Implicit". Yet, the preference for the implicit-duration model was equal or higher than the baseline in the first test, and not significantly different in the second, neither among native English speakers nor among all the subjects. Looking at the results, we notice that although the total preference differences with respect to the "baseline" are almost the same for all three approaches, the proportion of subjects who chose the "None" option is significantly lower when the duration is used implicitly. In other words, the implicit-duration model produces more audible differences. The explanation for this is that whereas the usage of the external duration in an explicit way affects only to the phone durations, its usage to select the models also affects the other speech components. Therefore, the poorer duration estimation of the implicit-duration model seems to get compensated for by a better modeling of the dependencies between the duration and the spectrum, logF0 or aperiodicity. This hypothesis also explains why the combination of the "Explicit" and "Implicit" approaches yields the highest preference with respect to the baseline.

5. Discussion

To study in greater detail the effect of the duration on the spectrum, aperiodicity and F0, we analyzed the usage of the duration questions in their decision trees. In HTS clustering, the order in which nodes are split depends on the global increment on

Table 4: *Usage of the duration question for state and stream*

Stream - state	#tree leaves	Avg. leave appearance	Avg. question appearance	Proportion
Cep1	620	397	428	107.8%
Cep2	666	420.5	386	91.8%
Cep3	841	533.7	423	79.3%
Cep4	682	436.6	394.7	90.4%
Cep5	607	396	398.7	100.6%
Ap1	847	518.9	397.7	76.7%
Ap2	1146	701.3	521.1	74.3%
Ap3	1448	887.2	572.9	64.6%
Ap4	1093	673.5	493.7	73.3%
Ap5	927	568.3	476.7	83.9%
logF01	920	575.3	480.6	83.4%
logF02	2151	1311.8	1074.2	81.9%
logF03	3152	1908.4	1483.8	77.8%
logF04	1805	1107.6	844.7	76.3%
logF05	1128	704.2	610.4	86.7%
dur	282	179.3	87.4	48.8%

maximum-likelihood that such splits provides. In other words, the earlier the average appearance in the tree of a set of questions the higher the log-likelihood increment they yielded, and thus the more important they are. To analyze the dependency of each factor with respect to the duration, we computed the average appearance position of the duration questions in the clustering trees, and compared it with the average appearance position of the tree leaf nodes. Table 4 shows the results of this analysis. As expected, the earliest average appearance of the duration questions is in the duration tree, around 50% of the leaf nodes. One result of this high position is that small errors in the input produced large errors in the output. In the other three streams, the duration questions appear much later. For the spectrum and aperiodicity trees, the average appearance of the duration questions is earlier in the central state of the phone and decreases toward the first and last state. This indicates that the duration of current and surrounding phones is less important at the transition between phones than in their central part, especially for the spectrum. The average appearance in the aperiodicity trees is higher than in the spectrum tree, which suggests that the duration has a stronger effect on the glottal source than on the vocal tract. The average appearance of the duration question in the logF0 trees does not decrease toward the model edges as sharply as in the aperiodicity and spectrum trees, probably due to the suprasegmental nature of logF0. On average, duration questions appear at around 80.1% of the way into the logF0 tree. This relatively high dependency can explain the audible prosody differences found between the stimuli synthesized with the baseline model and those synthesized with the implicit-duration model.

6. Conclusions

We have analyzed three possible ways to integrate an external duration model in an HMM-based synthesizer: the standard explicit definition of the phone duration, a novel implicit-duration approach that uses the external duration to select the synthesis model, and a combination of these two methods. A subjective evaluation comparing these approaches with a baseline HSMM-model showed an small but significant improvement in all three cases. However, the reason for the improvement is different for each approach. In the explicit approach, the improvement is the result of a better estimation of the phone duration by the external duration model. In the implicit-duration approach, the phone duration estimation is worse than the baseline, but the modeling of the dependencies between phone durations and the other speech components is better, especially for the aperiodicity and logF0. Such better modeling results in a more audible difference with respect to the baseline, which translates into ca. 20% less 'no-preference' choices. In the combined approach, the advantages of the other two are added. Consequently, it yields the highest preference with respect to the baseline.

7. References

[1] O. Goubanova and S.King, "Bayesian networks for phone duration prediction", Speech communication, v. 50 (4), pp. 301-311, April 2008

[2] K. Iwano, M. Yamada, T. Togawa and S. Furui, "Speech-rate-variable HMM-based Japanese TTS system" In Proc. IEEE Workshop on Speech Synthesis, 2002, Santa Monica, USA, pp. 219-222

[3] J.P.H. Van Santen, "Assignment of segmental duration in text-to-speech synthesis", Computer Speech and Language, v. 8, pp. 95-128, 1994

[4] K. Sreenivasa Rao and B. Yegnanarayana, "Modeling duration of syllables using neural networks" Computer Speech and Language, v. 21(2), pp. 282-295, 2007

[5] M.D. Riley, "Tree-based modeling for speech synthesis", In: G. Bailly, C.Beno it and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 265-273, 1992

[6] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features". Proc. ICASSP, 1995.

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Hidden Semi-Markov Model Based Speech Synthesis", in Proc. ICSLP, vol. II, 2004, pp. 1393-1396

[8] C. Hayashi, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view". Ann. the Institute of Statistical Mathematics Vol 3, no 2, pp.69-98, 1952.

[9] Z.H. Ling, Y.J. Wu, Y.P. Wang, L. Qin and R.H. Wang, "USTC System for Blizzard Challenge 2006 and improved HMM-based speech synthesis method" Proc. Blizzard Challenge workshop 2006, Pittsburgh, http://festvox.org/blizzard/bc2006/ustc_blizzard2006.pdf

[10] B. Gao, Y. Qian, Z. Wu, F.K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood" Proc. Interspeech 2008, Brisbane, Australia, pp. 2266-2269

[11] T. Masuko, "HMM-Based speech synthesis and its applications", Doctoral dissertation, Tokyo Institute of Technology, November 2002, pp. 34-35, http://www.kbys.ip.titech.ac.jp/masuko/masuko-doctor.pdf

[12] "HMM-based Speech Synthesis System (HTS)" Nagoya Institute of Technology, http://hts.sp.nitech.ac.jp/

[13] P.J.B. Jackson and C.H. Shadle, "Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech" in IEEE Transactions on Speech and Audio Processing, V.9 (7), pp. 713-726, October 2001

[14] Y. Stylianou, "Concatenative speech synthesis using a harmonic plus noise model", Third ESCA Speech Synthesis Workshop, pp. 261266, Nov. 1998.

[15] K. Tokuda,T. Masuko,N. Miyazaki and T. Kobayashi, "Multispace probability distribution HMM," in IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp. 455464, 2002.

[16] G. Webster and S. Buchholz, "Automatic feature selection from a large number of features for phone duration prediction" submitted Speech Prosody 2010, Chicago, USA.