UNSUPERVISED TRAINING OF PHONE DURATION AND ENERGY MODELS FOR TEXT-TO-SPEECH SYNTHESIS

Paul C. Bagshaw

France Télécom, CNET (Centre National d'Etudes des Télécommunications)

ABSTRACT

A new model of phone duration and energy is presented. These parameters are modelled in two stages. The first stage builds a statistics tree that contains phone duration and energy mean and standard deviation values at each node. The branches of the tree are characterised by a set of factors related to phonetic context. The second stage considers phone duration and energy to be modified by two syllable-level prosodic coefficients. The duration and energy of the phones of a syllable are influenced to differing degrees by these coefficients. Weights are associated with the different phone positions in a syllable. A simulated annealing technique is used to find the set of weights that allow the prosodic coefficients to be calculated for all syllables and, in turn, minimise the error in predicting the phone duration and energy during synthesis. They are predicted with a mean squared error of 15.4ms and 6.8dB respectively. During synthesis, the syllable-level prosodic coefficients are predicted by regression trees from linguistic information. Manual prosodic labelling is not required at any stage.

1. INTRODUCTION

The prediction of phone fundamental frequency (F0), duration and energy parameters is a crucial component of text-to-speech synthesis (TTS) systems. These parameters can be determined from discrete phonological labels, such as pitch accent and boundary tones, and break indices [10]. Such phonological labels must themselves be predicted from the results of linguistic analyses in a TTS system. There are two major weaknesses in this approach. Firstly, it is difficult to accurately predict the continuum of acoustic parameters from a small set of discrete labels. Secondly, systems that predict the phonological labels require voluminous training data to be manually annotated.

This paper deviates from the above approach and concentrates on phone duration and energy modelling (F0 modelling is left for a latter investigation). A technique is proposed in section 3 that enables two continuous coefficients to be automatically associated with each syllable in a corpus of phonemically transcribed speech — a syllable prominence coefficient, *p*, and a syllable lengthening coefficient, *l*. During synthesis, phone duration and energy parameters can be accuracy predicted from these coefficients. The prosodic coefficients, *p* and *l*, need to be predicted from discrete features (part-of-speech tags, form of syntactic parent, etc.) generated from linguistic analyses. This is done in section 4 by regression trees that have been adapted for multidimensional data. Learning *p* and *l* from the output of linguistic analyses is no more difficult than learning to predict discrete phonological labels. The technique is unsupervised in the sense that the two prosodic coefficients can be calculated without human intervention during both the analysis stage (from the acoustic signal) and the synthesis stage (from text). The tedious task of manually annotating a corpus with discrete prosodic labels can therefore be avoided.

2. DATABASE DESCRIPTION

The speech data used in this study consists of 460 'phonetically compact' sentences designed to provide as complete a coverage of phoneme pairs as possible [7], and read by a male speaker of British English (South-eastern accent). The orthography of the read text is used in conjunction with a lexicon of phonemic transcriptions to aid an HMM-based automatic phonemic segmentation [3] of the speech signals; the same lexicon is used in the speech synthesis system. The segmentation includes syllable and word boundaries, and lexical stress marks. The phonemic segmentation and word alignments are manually verified/corrected by a trained phonetician. The data constitutes 3509 words (5579 syllables).

The energy contour for a speech waveform (sampled at 16kHz) is calculated from 20ms frames at 5ms intervals. Each frame is passed through a Blackman-Harris window and the frequency bins of a power spectrum (512-point FFT) corresponding to the range 50Hz–2kHz are accumulated. These energy values are expressed in decibels with respect to the maximum frame energy in the utterance to form an utterance-normalised sonorant energy contour. The contour is processed by a five-frame median filter and five-frame Hann window smoother [9] in order to remove small perturbations which arise during frames of speech with low fundamental frequency. A phone-level energy contour is generated from the resultant (frame-level) low-band energy contour. Each phone is associated with the energy of one of the frames within the phone. If one or more local maxima exist then the value at the highest peak is associated with the phone. If a single minimum exists then its value is used. Otherwise, no local peak or valley can be located, the phone is associated with the energy value at its mid-point. The phone-level energy values and the duration of phones have a proven strong correlation with the perception of prominent syllables in English speech [1].

The text corresponding to each utterance is analysed by the linguistic processing modules of the *Anglovoc* speech synthesis system developed at France Télécom, CNET [4, 2]. The analysis assigns one of 36 possible part-of-speech tags to each word with an accuracy of 91.81% correct, and generates a complete syntactic parse with as many as 21.52% erroneous (crossing) brackets and 93.74% correctly labelled syntactic forms — results obtained by comparison with the Penn Treebank-II [8]. This data is aligned with the word-level segmentation.

3. PARAMETER ESTIMATION

3.1. The Model

Each syllable is assumed to be characterised by two continuous prosodic coefficients; a degree of syllable prominence, *p*, and lengthening, *l*. The duration and energy of each phone within a syllable need to be estimated during synthesis from these two variables. The estimated duration, $\hat{d}_i$, and energy, $\hat{e}_i$, of a phone are given by:

$$\hat{d}_i = \mu_{di} + (\omega'_{di} \cdot p + \omega''_{di} \cdot l) \cdot \sigma_{di} \quad (1)$$

$$\hat{e}_i = \mu_{ei} + (\omega'_{ei} \cdot p + \omega''_{ei} \cdot l) \cdot \sigma_{ei} \quad (2)$$

where μ_{di} and μ_{ei} are the mean duration and energy respectively for a phone of type *i*, σ_{di} and σ_{ei} are their corresponding standard deviations, and ω'_{di}, ω''_{di}, ω'_{ei} and ω''_{ei} are weights which modify the degree to which *p* and *l* affect the duration and energy for a phone of type *i*.

The means and standard deviations are calculated as a function of the phonetic context of a phone, as described in section 3.2. The manner in which *p* and *l* are derived is introduced in section 3.3. The weights are defined as a function of the position of the phone within a particular syllabic structure. The way in which their values are determined is given in section 3.4.

3.2. Phonetic Level

The mean and standard deviation values are determined from the database by partitioning phones into groups, whereby each group contains phones with similar phonetic contexts (independent of their prosodic context). Factors $\boldsymbol{P_{1\ldots4}}$, below, influence duration and energy at a phonetic level and are easy to obtain during synthesis. Their values define the type of phone, *i*.

- $\boldsymbol{P_1}$: is the phone label (phonemic class) — 47 categories.
- $\boldsymbol{P_2}$: if the phone is a consonant, specifies the position taken by the consonant in a cluster and the total number of consonants in the cluster (1/1, 1/2, 2/2, 1/3, 2/3, 3/3). If the phone is a vowel, specifies the number of consonants in the coda following the vowel (0, 1, 2, 3) — 10 categories.
- $\boldsymbol{P_3}$: if a consonant, specifies if the phone is in the syllable onset or coda. If a vowel (syllable nucleus), specifies if it is in an open syllable, a closed syllable followed by a 'lengthening' consonant (/l ɹ w y m n ŋ θ v ʒ/), or a closed syllable followed by a 'shortening' consonant — 5 categories.
- $\boldsymbol{P_4}$: says whether or not the phone is in a word-final syllable — 2 categories.

The mutual information (equation (3)) between each of these factors (assigned *x*) and the parameter duration or energy (assigned *y*) is calculated. In order to do this, it is necessary for the continuous parameter, duration or energy, to be quantised. A parameter is grouped into $\lfloor \log_2 N \rfloor + 1$ categories, each having the same interval, where *N* represents the number of observations.

$$I(x,y) = \sum_{i=1}^{X} \sum_{j=1}^{Y} \Pr(x_i, y_j) \cdot \log_2 \frac{\Pr(x_i, y_j)}{\Pr(x_i) \cdot \Pr(y_j)} \quad (3)$$

$\boldsymbol{I(x,y)}$	Duration	Energy
$\boldsymbol{P_1}$	0.417	0.829
$\boldsymbol{P_2}$	0.056	0.334
$\boldsymbol{P_3}$	0.052	0.305
$\boldsymbol{P_4}$	0.011	0.016

Table 1: Mutual information between the quantised parameters (duration and energy) and factors that influence them at a phonetic level.

The classification of a consonant as either a 'lengthening' or a 'shortening' consonant — for factor $\boldsymbol{P_3}$ — is made to maximise the sum mutual information between the factor and duration and energy; i.e. $I(\boldsymbol{P_3}, d) + I(\boldsymbol{P_3}, e)$.

The population of phones is recursively partitioned by each of the discrete phonetic factors, $\boldsymbol{P_{1\ldots4}}$, in decreasing order of mutual information (Table 1). The mean and standard derivation are calculated for the duration and energy of the phones in each group; but only if there are at least 20 phones in the group, otherwise the statistical estimators are not reliable. A tree-type structure is established, in which the root node gives the mean and standard deviation for all phones, and asks for the value of the factor $\boldsymbol{P_1}$. Each node in the second level of the tree gives the mean and standard deviation for each phonemic class, and requests the value of the next important factor (that which has the highest mutual information).

During synthesis, the type of a phone, *i*, is characterised by the values of the factors $\boldsymbol{P_{1\ldots4}}$. These values are used to traverse the tree of statistical estimators up to either a terminal node or a node where no child node exists for the given value of the factor under interrogation. In this way, values are retrieved for μ_{di}, σ_{di}, μ_{ei} and σ_{ei}. If a phone is encountered during synthesis that has a phonetic context underrepresented in tree training data, then the retrieved statistical estimators will be the most befitting.

3.3. Prosodic Level

The prosodic coefficients for prominence, *p*, and lengthening, *l*, need to be determined for each syllable such that the error of estimating the duration and energy for each of its component phones is minimised. The objective function of the optimisation *F* is defined as the mean squared error of the duration and energy:

$$\min_{p,l} \arg F(l,p) = (1-\beta) \cdot \sum_{i \in syllable} (\mu_{di} + l \cdot \sigma'_{di} + p \cdot \sigma''_{di} - d_i)^2 + \beta \cdot \sum_{i \in syllable} (\mu_{ei} + l \cdot \sigma'_{ei} + p \cdot \sigma''_{ei} - e_i)^2 \quad (4)$$

where $\sigma'_{di} = \omega'_{di} \cdot \sigma_{di}$, $\sigma''_{di} = \omega''_{di} \cdot \sigma_{di}$, $\sigma'_{ei} = \omega'_{ei} \cdot \sigma_{ei}$ and $\sigma''_{ei} = \omega''_{ei} \cdot \sigma_{ei}$.

The coefficient $\beta \in]0,1[$ is used to place a bias on the duration or energy. The objective function is solved by setting the first partial derivatives to zero. Note that the second partial derivatives are always greater than zero, therefore a minimum is located when the first partial derivatives are zero. The solution is given by:

$$\begin{pmatrix} p \\ l \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} \text{ if } \Delta \neq 0 \tag{5.1}$$

where

$$a_{11} = \sum_{i \in syllable} \left[(1-\beta) \cdot {\sigma''_{di}}^2 + \beta \cdot {\sigma''_{ei}}^2 \right] \tag{5.2}$$

$$a_{22} = \sum_{i \in syllable} \left[(1-\beta) \cdot {\sigma'_{di}}^2 + \beta \cdot {\sigma'_{ei}}^2 \right] \tag{5.3}$$

$$a_{12} = a_{21} = \sum_{i \in syllable} \left[(1-\beta) \cdot \sigma'_{di} \cdot \sigma''_{di} + \beta \cdot \sigma'_{ei} \cdot \sigma''_{ei} \right] \tag{5.4}$$

$$A = \sum_{i \in syllable} \left[(1-\beta) \cdot \sigma''_{di} \cdot (d_i - \mu_{di}) + \beta \cdot \sigma''_{ei} \cdot (e_i - \mu_{ei}) \right] \tag{5.5}$$

$$B = \sum_{i \in syllable} \left[(1-\beta) \cdot \sigma'_{di} \cdot (d_i - \mu_{di}) + \beta \cdot \sigma'_{ei} \cdot (e_i - \mu_{ei}) \right] \tag{5.6}$$

$$\Delta = a_{11} a_{22} - a_{12} a_{21} \tag{5.7}$$

If the determiner $\Delta = 0$, such as when $\sigma'_{di} = \sigma''_{di} \forall i$ and $\sigma'_{ei} = \sigma''_{ei} \forall i$, then it is necessary to solve the simplified case:

$$\begin{aligned} \min_k \arg F(k) = (1-\beta) \cdot & \sum_{i \in syllable} (\mu_{di} + k \cdot \sigma_{di} - d_i)^2 \\ & + \beta \cdot \sum_{i \in syllable} (\mu_{ei} + k \cdot \sigma_{ei} - e_i)^2 \end{aligned} \tag{6}$$

The solution that satisfies the first derivative equal to zero is given by:

$$k = \frac{\sum_{i \in syllable} \left[(1-\beta) \cdot \sigma_{di} \cdot (d_i - \mu_{di}) + \beta \cdot \sigma_{ei} \cdot (e_i - \mu_{ei}) \right]}{\sum_{i \in syllable} \left[(1-\beta) \cdot {\sigma_{di}}^2 + \beta \cdot {\sigma_{ei}}^2 \right]} \tag{7}$$

In this simplified case, let $p = k$, $l = 0$ if the syllable is not word-final; otherwise assume that the prosodic modification of the syllable is shared equally between the prominence and lengthening coefficients, i.e. $p = l = k/2$.

3.4. Weights

A phone has one of five types given by the feature $\boldsymbol{P_3}$ described in section 3.2. Thus, a syllable has one of the following six structures: nucleus, onset-nucleus, long-coda, short-coda, onset-long-coda, or onset-short-coda. Each phone therefore takes one of 13 possible positions within a syllable; the onset of a syllable containing onset-nucleus, the nucleus of a syllable containing onset-nucleus, etc. A phone exists in either a word-final or a non-word-final syllable, as given by feature $\boldsymbol{P_4}$. There are hence a total of 13×2 possible phone types, *i*, for each of ω'_{di}, ω''_{di}, ω'_{ei} and ω''_{ei} (104 weights).

A number of assumptions and constraints are imposed onto the weights. First, it is assumed that syllables not in word-final position are not lengthened; hence:

$$\omega''_{di} = 0, \ \omega''_{ei} = 0 \tag{8}$$

for all phone types not within a word-final syllable.

Second, if one phone in a particular syllable is more/less susceptible to *p* or *l* than another phone in the same syllable, then it is assumed to have taken/given the balance of the influence made on the duration or energy by the prosodic coefficient from/to the other phone. Hence the weights in a particular syllable structure in a given word position are constrained such that their average is one.

For a syllable in either word-final and non-word-final position,

$$\frac{1}{|S|} \cdot \sum_{i \in S} \omega'_{di} = 1, \ \frac{1}{|S|} \cdot \sum_{i \in S} \omega'_{ei} = 1 \tag{9.1}$$

For a syllable in word-final position only,

$$\frac{1}{|S|} \cdot \sum_{i \in S} \omega''_{di} = 1, \ \frac{1}{|S|} \cdot \sum_{i \in S} \omega''_{ei} = 1 \tag{9.2}$$

where $|S| = 1$ if *S* is a syllable structure with no onset or coda (*i* = nucleus only); $|S| = 2$ if *S* represents a syllable structure either of the form onset-nucleus, long-coda or short-coda; and $|S| = 3$ if syllable structure *S* is onset-long-coda or onset-short-coda. This second constraint is essential if values of syllable prominence and lengthening derived (during training) from one type of syllable structure are to be applied (during synthesis) to another type.

After taking equations (8), (9.1) & (9.2) into account, there are 42 weights remaining. The values for these weights are determined by a process of adaptive simulated annealing (ASA) [6]. It is necessary to impose a range constraint on the weights for the ASA process. The third constraint is:

$$\omega \in [0.0, 2.0] \tag{10}$$

ASA makes an initial random hypothesis of the values of these 42 weights. The first two constraints are used to determine values for the remaining required weights. Values of *p* and *l* are then calculated for every syllable in the database by using the method described in section 3.3. The duration and energy for every phone in every syllable are then calculated from equations (1) and (2). ASA adapts the values of the weights, considering the third constraint, in such a way as to minimise the cost function, *C*, given by:

$$C(\Omega) = (1-\beta) \cdot \sum_i (\hat{d}_i - d_i)^2 + \beta \cdot \sum_i (\hat{e}_i - e_i)^2 \tag{11}$$

where Ω represents the set of weights. Note that equation (11) is equivalent to equation (4), except that in this case, the errors are summed across all phones in the database.

3.5. Objective Evaluation of Model

The statistical estimators, prosodic coefficients and weights are determined for the database described in section 2. The biasing coefficient is arbitrarily set, $\beta = 0.2$. The root mean squared (r.m.s.) error in estimating the duration of phones is 15.4ms. Phone energy is estimated with an r.m.s. error of 6.8dB. These results show that phone duration and energy can be reliably estimated from exact values for the syllable prominence and lengthening coefficients.

If all weights are set to 1.0 then the model reduces to a simple system (equation (7)) with a single coefficient of syllable elasticity [5]. In this case, the duration error rises to 21.2ms and the energy error rises to 7.5dB.

4. PREDICTION OF SYLLABLE COEFFICIENTS FROM TEXT

The continuous prosodic coefficients *p* and *l* must be predicted during synthesis from linguistic information that is derived from the text and that is discrete. Two regression trees are leant with a splitting criterion based on multivariate analysis of variance, adapted from the *RPART* program [11]. One tree predicts only a value of *p* for syllables not in a word-final position $(l = 0)$, and the other predicts a duple (p, l) for word-final syllables.

The first tree contains 121 splits and predicts *p* with a mean squared error of 0.318. Its top nodes interrogate the lexical stress of the current syllable, a complexity index of the syntax structure and the part-of-speech tag for the two previous words and the following word. The second tree contains 185 splits and predicts *p* and *l* with mean squared errors of 0.616 and 0.622 respectively. Its top nodes interrogate the number of words before the next punctuation mark and the part-of-speech tag for the current word.

Boundary tone and pitch accent labels can also be predicted from the linguistic information. A classification tree may be trained to perform this task if hand-labelled data is available. When the estimated boundary tone and estimated pitch accent labels are included amongst the input parameters for the regression trees, the accuracy of the predicted values for *p* and *l* is not improved.

5. CONCLUSION

Continuous prosodic coefficients can be automatically calculated for all syllables in a corpus. These coefficients can be used to estimate phone duration and energy during speech synthesis. Their values can also be predicted by regression trees, which take linguistic information as input. There is no need for boundary tone and pitch accent labels to be amongst the input parameters. The entire system can therefore be trained without any hand-labelling of prosodic events.

6. REFERENCES

1. Bagshaw, P.C. (1994) *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching*. Ph.D. thesis. University of Edinburgh, Scotland, U.K.

2. Bagshaw, P.C. (1997) *Anglovoc-II : Traitements Linguistiques et Prosodiques pour la Synthèse de la Parole en Anglais Américain et Britannique*, Note Technique NT/DIH/RCP/00112, Centre National d'Etudes des Télécommunications: Lannion, France.

3. Boëffard, O. (1993) *Segmentation Automatique d'Unités Acoustiques pour la Synthèse de la Parole*. Ph.D. thesis. Université de Rennes I, France.

4. Boëffard, O., Bigorgne, D., Cherbonnel, B., Emerard, F., Roussarie, L., Bagshaw, P., Conkie, A., Ennilo, M. & Traber, C. (1996) Utilisation de techniques d'apprentissage automatique pour les traitements linguistiques et prosodiques en synthèse de la parole : quelques résultats en Anglais, Allemand et Français. *Actes des XXIes Journées d'Etude sur la Parole*, Avignon, France, 10–14 June, 383–386.

5. Campbell, W.N. & Isard, S.D. (1991) Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

6. Ingber, L. (1993) Adaptive Simulated Annealing (ASA). Global optimization C-code, Lester Ingber Research, Chicago, IL. [http://www.ingber.com/#ASA-CODE].

7. Lamel, L.F., Kassel, R.H. & Senett, S. (1986) Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. DARPA Speech Recognition Workshop, Pato Alto, California, 19–20 Feb. 1986.* (Baumann, L.S. ed.), Science Applications International Corporation: McLean, Virginia. 100–109.

8. Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**, 313–330.

9. Rabiner, L.R., Sambur, M.R. & Schmidt, C.E. (1975) Applications of non-linear smoothing algorithms to speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, **23**(6), 552–557.

10. Silverman, K.E.A., Beckman, M.E., Pitrelli, J. Ostendorf, M., Wightman, C.W., Price, P.J. Pierrehumbert, J.B. & Hirschberg, J. (1992) TOBI: A standard for labelling English prosody. *Proc. International Conference on Spoken Language Processing*, Banff, Canada, **2**, 867–870.

11. Therneau, T.M. & Atkinson, E.J. (1997). An introduction to recursive partitioning using the RPART routines. Mayo Foundation: Rochester, MN. [http://lib.stat.cmu.edu/general/rpart].