Optimal Footprint for Prosodic Modelling

Eric Keller & Brigitte Zellner Keller

LAIP, IMM - Lettres
University of Lausanne, Switzerland
{eric.keller}{brigitte.zellnerkeller}@imm.unil.ch, http://www.unil.ch/imm/docs/LAIP/LAIP.html

Abstract

We examined the extent of material required to build prosodic models for duration, fundamental frequency and intensity. 50 multiple linear regression models were built for two MARSEC speakers on the basis of 70 utterances (7'522 and 7'643 segments). Models based on 8 and 20 utterances showed closeness of fits comparable to those reported by other researchers for much larger corpora. Little systematic improvement was seen beyond 20 utterances. A predictor ranking procedure advantageously replaced the more commonly used regression trees. Results suggest that a series of well-adapted small-footprint models provide more accurate information about the individual use of prosody in specific speech situations than a single model based on abundant data[1].

1. Introduction

Numeric prosodic models are generally used to classify and synthesize the fine details of language behaviour and they are frequently used to simulate speech behaviour over time. In such models, the size of the underlying database is an important issue. A model is *underspecified* if it is based on too little material. In this case the abstracted "rules" of the model do not generalize well to new material. Alternatively, a model can be *overspecified* if it is based on too much material and cannot reflect local variations in speech or speaker style. An optimal compromise between generalizability and local accuracy must thus be found.

The issue is complicated by the fact that neither the prosodic parameters nor the segments or syllables to which they apply are evenly distributed in normal speech (the "sparse-data problem" [5], [17], [13]). The optimal compromise for a frequent parameter (e.g., *f0*) is thus not necessarily the same as that for a rarely-occurring parameter (e.g., a turn-taking signal). In the present study, we address this problem in two ways. First, we concentrate on the requirements for building models for the central and coherent three-parameter set of segmental duration (*t*), fundamental frequency (*f0*) and intensity (*I*). In the case of *t* and *I,* a parameter specification is required for every segment, and in the case of *f0*, a specification is required for every voiced segment. This assures a high frequency of occurrence and high local accuracy for the most common segments of the language. Second, the effects of underspecifying parameters for rarely occurring segments (e.g., /Z/ as in "measure" in English) are examined, as when predictors for infrequent values are interpolated or simply set to the mean value[2].

To sum up, the desired compromise must satisfy three criteria: *generalizability*, *local accuracy* and *optimal coverage*. An approximation can be found by examining the evolution of the following three measures in models generated from databases of increasing size: *closeness of fit of the model*, *capacity of prediction to novel material* and *coverage of segments*. Our results indicate that at least for certain speakers reading prepared materials, an optimal compromise can be based on databases that are much smaller than those that are commonly used in numeric *t-f0-I* modelling.

2. Method

2.1. Data Preparation

Corpus. Numeric prosodic models were built for two speakers with extensive materials in the Machine-Readable Spoken English Corpus (MARSEC[3][10], [11]). Both are practiced adult speakers representative of the received-pronunciation (*i.e.*, the prestige/regionally neutral accent) form of British English (RP). Materials were (1) one and a half news bulletins spoken by Brian Perkins (BP) a BBC-4 newscaster (the initial part of MARSEC section B, 70 utterances, 7'522 segments), and (2) a portion of the 1985 Reith Lecture presented by the economist David Henderson (DH) (the initial part of section C, 70 utterances, 7'643 segments). Punctuation was inserted into MARSEC's text transcriptions at major tone group ("utterance") boundaries. With a few exceptions motivated by semantics and syntax, such markers were recoded into appropriate periods, question/exclamation marks, and colons. Commas were set according to English punctuation rules. The acoustic material from the two speakers was phone-level segmented using a computer aided-technique of placing marks that are subsequently adjusted manually.

Independent variables (IVs). On the basis of the written and punctuated texts adjusted for hesitations and speech errors, 19 linguistic predictors were obtained. Most of these have been shown to be of statistical relevance for the prediction of duration ([2], [4], [5], [6], [14], [15]):

- **Positional information** (5)[4]: segment position in syllable, syllable position in word, word position in minor phrase, minor phrase position in major phrase, major phrase position in utterance). "Major phrases" were text segments delimited by sentence markers and by commas. "Minor phrases" were delimited by punctuation marks as well as boundaries between lexical and grammatical words, as defined below. Such position indicators have been shown to be of relevance for the prediction of duration for French [12], [6], [18], [19], and have

[1] This is a short description of a larger study [8]. Please refer to the main paper for details of methodology, illustrations and discussion.
[2] The sparse-data problem can be circumvented by using phonetically-balanced reading material. However, that introduces other problems, such as inapplicability of the approach to spontaneous speech, a time-consuming preparatory phase, and the common observation that desired phonetic features are either not pronounced by speakers or were forgotten or not foreseen in the constitution of the database.
[3] MARSEC is available from www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec.
[4] In parentheses: number of IVs of this type.

also been successfully applied to German in our laboratory [16], as well as in various other languages (e.g., [1], [17], [3]).

- **Quantitative information** (5): number of segments in syllable, syllables in word, words in minor phrase, minor phrases in major phrase, major phrases in utterance.
- **Segment boundaries** (1): no boundary, syllable, lexical, minor phrase, major phrase, semicolon, period, question mark, exclamation mark and colon.
- **Phonemic segment identity and lexical stress** (6): identity of preceding, current, and succeeding segments, lexical stress of preceding, current and succeeding segments.
- **Part-of-speech [POS] membership** (2): the full POS classification according to the CUVOALD electronic dictionary, and a simplified POS, where the CUVOALD classifications G Anomalous verb (e.g., auxiliary), Q Pronoun, R Definite article, S Indefinite article, T Preposition, U Prefix, and V Conjunction were set to "grammatical word" and the rest of the Cuvoald POS classifications (nouns, full verbs and adjectives) were considered "lexical words".

Dependent variables (DVs). Prior to the extraction of the dependent variables, signals were denoised, enhanced, and converted to 16 kHz with Pristine Sound[1] after low-pass filtering. DVs were as follows: (a) segmental *t*, (b) *f0* extracted at 500 Hz by Praat's[2] AC routine with default settings, except for a pitch ceiling of 250 Hz, and (c) *I* extracted at 500 Hz by Praat's intensity routine, with minimum pitch set to 75 Hz. *F0* and *I* curves were spot-checked for abnormalities. For each segment, 10 equally-spaced *f0* and *I* values were derived from the extractions and mean values were stored for each segment. For ease of data manipulation, Praat's interpolated *f0* values were used for unvoiced segments, although physically no voicing occurred, and logically no *f0* value could apply. Although this manner of proceeding bore the risk of depressing *f0* prediction values, this was considered acceptable, since (a) this study focused on relative values and a depression in absolute predictions had no bearing on the interpretation of the results, (b) prediction was unaffected for segments for which voicing is audible, and (c) the ultimate reduction in prediction proved to be minimal (see below). Pauses were not modelled in this study.

IV and DV information was aligned semi-automatically. For example, CUVOALD, used for generating the IVs, provides rhotic transcriptions (/kA:r/ for "car"), while the two speakers generally use non-rhotic pronunciations in the DV segmentations (e.g. /kA:/ for "car"). Since some r's were pronounced, manual adjustments were needed to remove excess IV segment information. Similarly, some optional fast-speech rules, such as the deletion of schwa in "fin**a**lly" or of the second /t/ in "contract **t**alks", were performed manually. After each modification, IV positional and segment context was automatically updated to reflect the segments' new status.

2.2. Statistical Processing

Before calculating multiple regressions on the IV-DV relationships, the following procedures were applied:

(1) *DV Linearization*. Various power transformations were applied to the three DVs (log, square root, raw data, and power of 2), and the best linearizing transformation was chosen as judged by proximity to zero on a calculation of skewness. For both speakers, the best transformations were square root for segment *t*, log for *f0*, and power of 2 for *I*.

(2) *IV Ranking*. For improved predictive accuracy, nominal IV distinctions were converted into ordinal IVs by placing average linearized DV values into IV predictor cells. E.g., the short vowel /I/ has an average duration of 58 ms for DH. Since this is a duration, 0.2409 (square root of 0.058 s) was placed in the duration-predicting IV cell for this segment. If a numeric IV value was absent from the corpus (e.g., if stress 0 and stress 2 were attested, but stress 1 was not), the interpolated value was taken. If a nominal IV was unattested but exists in the language (such as /Z/), the DV's mean value was used. All IVs were re-coded in this manner, separately for each IV-DV relationship.

This procedure permits ordinal predictors to be used instead of less powerful nominal or categorical predictors. To illustrate with an example from our previous studies, the best predictor of segmental *t* in French and German had been the segment's phonemic identity (fricatives tend to be long, unaspirated stops tend to be short, etc.). Since we could not discover any articulatory or acoustic logic for ranking segment identities along the duration scale, our initial option had been to treat them as nominal elements. This was not very appealing, since one of the applicable mechanisms (regression with nominal values, as found in most statistics packages) was much less powerful than a regression with ordinal or scale values, and another (frequently-employed regression trees, e.g., [14][9]) required quite a bit more data, due to the sparse-data problem. A multiple regression for the modelling of duration based on simple nominal values would explain only about 10-20% of variance, while regressions operating on DV-ranked data generally explain more than 50%, and sometimes as much as 79% of variance [19].

In our previous work, segment IVs were thus grouped into "duration groups" [6]. Zellner [19] proposed a grouping on the basis of purely quantitative criteria, a solution subsequently applied to German [16]. Although generally successful, this grouping of values also led to a certain loss of predictive power. Furthermore, a given grouping may make good predictions for one IV-DV relationship (e.g., the current segment identity – duration relation) but not for another (e.g., the relation between the identity of the preceding segment and current duration). The present study preserves the greatest possible predictive power by abandoning categorisation, by treating each IV-DV relationship separately, and by letting the data itself determine ordinal ranking of IVs instead of imposing groups preconceived on theoretical grounds. Also, the approach is generalized here to *f0* and *I*.

(3) *Outliers,* defined as data points lying two RMSEs or more above or below the predicted value, were examined and appropriate modifications were applied. The prediction model was recalculated after removal/modification of outliers.

(4) *Variable Pruning.* After the calculation of an initial set of regressions for the two speakers, IVs showing multicollinearity were removed with the aid of the SSPS stepwise procedure. This assured orthogonality in predictor variables that are significantly related to DVs ($p<0.05$), thus stabilizing the regression coefficients.

2.3. Evaluation and Precision

(1) Statistical Evaluation. Model performance was measured by the correlation coefficient *r*, the root mean-squared error (RMSE), and by *bias*, i.e., the degree to which the model under- or over-predicts target values. RMSE compensates for differences in corpus size and takes into account bias, which is left aside by correlation-based

[1] Available at http://www.accuratesound.net.
[2] Available for free at http://www.praat.org.

statistics. The examination of bias (a component of RMSE) can point up structural deficiencies in the model.

(2) Synthesis Evaluation. A *Hello World* message and the British version of the *Northwind Passage* were synthesized on the basis of completed models using predictions for segmental *t* and *f0* (not *I*). Four *f0* values per segment were generated, and a cubic spline function was used to smooth the *f0* curve. Pauses were simulated by static rules (end of major breaks: 100 ms, utterance final: 600 ms). Segmental *t*'s were linearly lengthened by 5% to facilitate perception and judgement. Outputs were created with Mbrola and the "en1" diphone database[1], and were denoised and filter band-enhanced with Pristine Sound before being converted to MP3. Sounds are available on the CD-ROM and at our website[2].

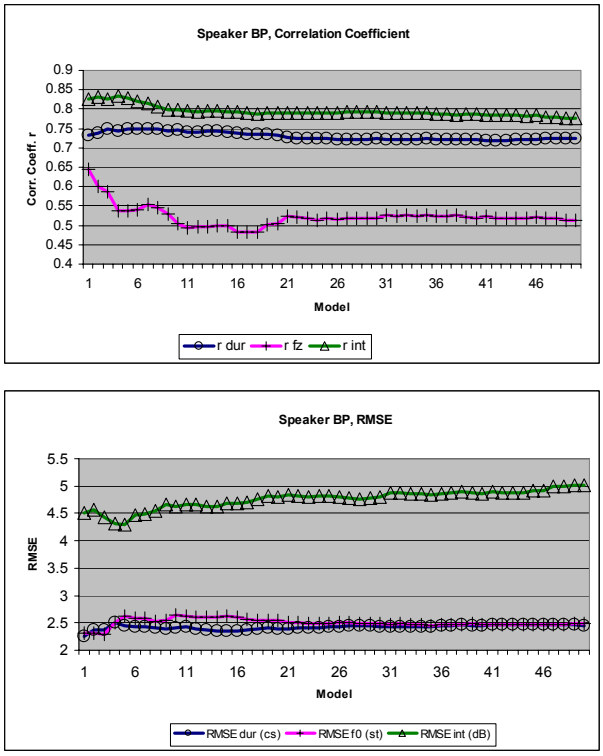

Figure 1. *Correlations and RMSEs for 50 models of increasing domain, where model 1 is based on utterances 1...2, model 2 on utterances 1...3, and model 50 is based on utterances 1...51 (Subject BP). The stability estimates are similar for the two speakers, even though the model fits for BP are better than those for DH.*

(3) Precision. To assure requisite control over precision in the model-building process, human intervention was applied at five points: (1) identification of predictor variables, (2) segmentation of the acoustic signal, (3) alignment between the dictionary-and-rule-based and the actual segment sequence, and (5) stepwise selection of IV parameters.

3. Results

3.1. Model Fit for the Original Data

50 models per DV were calculated over domains of increasing size, ranging from a model based two contiguous utterances (1 and 2), through to the last model, calculated on the basis of 51 utterances. Models were calculated separately for the two speakers' *t*, *f0* and *I*. Correlation coefficients and RMSEs show that the models stabilize rapidly beyond the fifth utterance, to attain approximate stability with the 20th utterance (Fig. 1). Estimates are similar for the two speakers, although model fits for BP are better than for DH. Measures for model 20 were as follows: (Speaker BP, N=1962) *t*, r = .735, p<.001, RMSE = 24 ms, bias = 2.3 ms; *f0*, r = .502, p<.001, RMSE = 2.55 semitones (16.7 Hz), bias = 1.24 Hz; *I*, r = .790, p<0.001, RMSE = 4.82 dB, bias = -0.17 dB. (Speaker DH, N=2331) *t*, r = .686, p<.001, RMSE = 28 ms, bias = 2.7 ms; *f0*, r = .294, p<.001, RMSE = 3.76 semitones (25.4 Hz), bias = 2.65 Hz; *I*, r = .721, p<0.001, RMSE = 5.56 dB, bias = -0.23 dB.

Since the *f0* model predicts values for both voiced and unvoiced segments even though unvoiced data were interpolated, a separate analysis was performed just for the voiced segments of model 20. It showed few if any improvements: correlations *r*'s for voiced segments (N=602) were .549 for BP (marginally better) and .294 for DH (N=742, unchanged), and RMSEs were 2.47 semitones (15.7 Hz) and 3.6 semitones (24.2 Hz) respectively, both marginally better.

The relationship between bias and RMSE was examined to assess models' over- or underestimation. Strong correlations between RMSE and bias (*t*: r=.906, *f0*: r=.988, *I*: r=-.904) indicated that bias accounts for nearly all of RMSE variation. However, percentages of RMSE accounted for by the bias component were not excessive (means of bias as a percent of RMSE for 50 models and the two speakers: *t*: 9.8% and 10.0%, *f0*: 7.3% and 10.2%, *I*: 3.5% and 4.0%). Models tended to underestimate duration and fundamental frequency, and overestimate intensity. This can in part be related to the physical and physiological nature of the scales involved. In normal speech, *t* is open-ended at the top but close-ended at the bottom, *f0* shows more range above median activity than below, and *I* shows the inverse tendency. This is reflected in the relationship between the mean and the median, where BP, for example shows a median below the mean for *t* (median: 48.6 ms, mean: 50.0 ms) and for *f0* (median: 100.7 Hz, mean: 102.2 Hz), but a median above the mean for *I* (median: 72.7 dB, mean: 71.8 dB). Since RMSE variation reflects primarily bias and captures small over- and underestimations of the modelled data, the variations shown here probably correspond to the presence of excessive values in the data (confirmed by a visual inspection of raw and bias values [8]).

As a final observation, *t* and *f0* values for model 7, based on 8 utterances and 779 (BP) / 978 (DH) segments, showed especially encouraging performance. Also a hint of a "counter-effect" is visible around models 17 and 13 respectively for those two parameters. A good short-term fit might thus be based on just 7-8 contiguous utterances. In contrast to the stable long-term model visible for utterance pools of 20+ utterances, the short-term model could well represent an optimum for capturing changing prosodic trends. This was borne out by sliding-window model fits for 70 sentences [8]. In short-term models based on a sliding 8-utterance window and longer-term models based on a sliding 20-utterance window, 8-utterance models tended to fit the data somewhat better than the 20-utterance models, both in the sense of reflecting more local variation and by reducing the error term somewhat (exception: *f0*).

[1] Mbrola: http://tcts.fpms.ac.be/synthesis/mbrola.html. "en1" by Alan Black, Paul Taylor, Roger Burroughs, Alistair Conkie, and Sue Fitt.

[2] http://www.unil.ch/imm/docs/LAIP/LAIPTTS_pros_footprint.htm.

3.2. Modelling Novel Data

The models' ability to generalize to novel speech material was examined by training 8- and 20-utterance models on the two speaker's initial utterances in the corpus and by applying predictions to utterances numbered 21-70. Correlations and RMSEs were calculated between predicted and measured values obtained from the original files [8].

Model fits for the novel data were somewhat weaker than for the original data and were quite variable, but still in the useful range (r's for t .63-.73, for $f0$ .35-.55 (BP) / .15-.45 (DH), and for I .66-.80). RMSEs were much less correlated with bias (r_t = .460, r_{f0} = .415, r_I = .049, average of two model types and both subjects), suggesting a different source of variation (e.g., variation in data set concordance). In line with BP's better-fitting models for the original data, BP's models had generally stronger predictive power than models based on DH's material. Models based on twenty utterances furnished only marginally better predictions than those based on eight utterances (an average of 1.3% improvement for correlations and an average of 0.9% improvement for RMSEs).

3.3. Segment Coverage

Missing data were calculated for the Northwind Passage on the basis of 8- and 20-utterance models. Interpolated missing numeric IVs (e.g., stress levels) accounted for 5% (BP) and 3.5% (DH) in the 8-utterance models, and for 1.4% and 2.1%, respectively, of the 20-utterance models. No average values needed to be substituted for missing nominal IVs (e.g., segments). It may be concluded that missing data did not play a major role in the present data set.

4. Conclusion

In conclusion, statistical and perceptual evaluations of *t-f0-I* prosodic models for news and lecture material suggest that an 8-utterance (<1000 segment) model can probably furnish useful short-term information, while a still very compact 20-utterance model can provide a stable estimate of longer-term behaviour. Synthesis results suggested that (1) there are few differences between models based on eight and twenty utterances, with the exception of *f0* prediction, (2) the better-fitting models based on BP's speech did appear to sound a bit better than those based on DH's speech material, particularly with respect to *f0*, and (3) some of the stylistic differences between BP's and DH's prosody could be identified in the synthesis by listeners familiar with the speakers' speech patterns, notably DH's considerable *f0* modulations.

RMSEs for duration in our two speakers ranged from about 21-29 ms. Klabbers [9], p. 70, using classification trees established on the basis of phonetic principles, reported duration RMSEs ranging from 19 to 27 ms on the basis of much more data (Dutch RMSE 27 ms, 12'948 segments, German RMSE 19 ms, 24'240 segments, and French RMSE 25 ms, 7'143 segments). We saw that beyond the 20th utterance, the variable of greatest impact was not the size of the data set, but the selection of the speaker (e.g., mean duration $RMSE_{n=8}$ for BP: 24.0 ms, DH: 26.5 ms).

Further improvements to this class of model can be expected from (1) the definition of additional and stronger IVs, (2) the modelling of interactions, (3) improved handling for missing data, and (4) using non-linear methods (such as neural nets) for prosodic parameters with strong non-linear components (such as *f0*) [8].

5. Acknowledgements

Many thanks go to Mr. L. Wiget for segmentation and development of text analysis routines. This research is supported by a Swiss OFES grant under COST 277.

6. References

[1] Campbell, W.N. 1992. Syllable-based segmental duration. In G. Bailly & C. Benoit (Eds.), *Talking Machines. Theories, Models, and Designs*. Elsevier Science Publishers, 211-224.

[2] Fant, G.; Kruckenberg, A.; Nord, L. 1991. Durational correlates of stress in Swedish, French and English. *Journal of Phonetics.* 19, 351-365.

[3] Febrer, A.; Padrell, J.; & Bonafonte, A. 1998. Modeling phone duration: Application to Catalan TTS. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 43-46.

[4] Huber, K. 1991. *Messung und Modellierung der Segmentdauer für die Synthese deutscher Lautsprache.* Diss. Nr. 9535, Institut für Elektronik, ETH Zürich.

[5] Keller, E. 2002. Towards greater naturalness: Future directions of research in speech synthesis. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in Speech Synthesis*. Wiley & Sons, 3-17.

[6] Keller, E.; Zellner, B. 1995. A statistical timing model for French. *XIIIth International Congress of Phonetic Sciences,* 3. Stockholm, 302-305.

[7] Keller, E.; Zellner, B. 1996. A timing model for fast French. *York Papers in Linguistics,* 17, 53-75. University of York.

[8] Keller, E.; Zellner Keller, B. (in press) How Much Prosody Can You Learn from Twenty Utterances? *Linguistik-online* (http://www.unil.ch/imm/docs/LAIP/LAIPTTS_pros_footprint.htm).

[9] Klabbers, E. 2000. *Segmental and Prosodic Improvements to Speech Generation.* CIP-DATA Library, T.U. Eindhoven.

[10] Knowles, G.; Wichmann, A.; Alderson, P. 1996. *Working with Speech.* Addison Wesley Longman.

[11] Knowles, G.; Williams, B.; Taylor, L. 1996. *A Corpus of Formal British English Speech*. Addison Wesley Longman.

[12] Malfrère, F.; Dutoit, T.; Mertens, P. 1998. Automatic prosody generation using suprasegmental unit selection. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 323-328.

[13] Möbius, B. 2001. Rare events and closed domains: Two delicate concepts in speech synthesis. *Proceedings 4th ISCA Workshop on Speech Synthesis.* Perthshire, Scotland.

[14] Riedi, M. 1998. *Controlling Segmental Duration in Speech Synthesis Systems*. PhD thesis, No. 12487, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 26, ISBN 3-906469-05-0), February.

[15] Riley, M. 1992. Tree-based modelling of segmental durations. In G. Bailly et al., (Eds.). *Talking Machines: Theories, Models, and Designs*. Elsevier, 265 - 273.

[16] Siebenhaar, B.; Zellner Keller, B.; Keller, E. 2002. Phonetic and timing considerations in a Swiss High German TTS system. in Keller, E., Bailly, G., Monaghan, A., Terken, J. & Huckvale, M. (Eds.) *Improvements in Speech Synthesis*. Chichester: John Wiley, 165-175.

[17] Venditti, J.J.; van Santen, J.P.H. 1998. Modeling segmental durations for Japanese text-to-speech synthesis. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 31-36.

[18] Zellner, B. 1996. Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée,* 1, 7-23.

[19] Zellner, B. 1998. *Caractérisation et Prédiction du Débit de Parole en Français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.