Abstract—Many phonetic and phonology domain research papers analyzed segmental duration: what factors and interactions between factors determine their duration. Their results often play an important role in Language Technology applications, for example TTS (text-to-speech synthesis), ASR (automatic speech recognition) widely used in infocommunication. Speech sound duration depends on various factors such as phonetic quality, phonological context, phonological position in the word or in the utterance, speech style, etc. We intended to automatically predict vowel duration in spontaneous speech based on three methods. (i) A classification/regression tree (CART) using some characteristic features of the vowel quality and context. (ii) The same features and feedforward neural network (FFNN) were used to model vowel duration. (iii) In the third method FFNN was used to predict vowel duration using the combination of characteristic features and spectral features. Empirical durational data were obtained by measuring vowel durations as attested in over 110 minutes of a large Hungarian spontaneous speech data base (BEA). Using CART there was a poor correlation (0.57) between measured and predicted vowel duration, with average RMSE (root mean square error) of approximately 33 ms. When using FFNN the results were slightly better: the correlation between the target and predicted vowel duration was 0.62 while RMSE was about 29 ms. When the combined features were used the results were even better: the correlation between the target and predicted vowel duration was 0.79 while RMSE was 25 ms. The results obtained for Hungarian support the complexity of features affecting vowel duration, on the one hand, while on the other they indicate the temporal complexity of segmental level of spontaneous speech, as has already been reported for Lithuanian, Czech, Hindi, Telugu and Korean.

I. INTRODUCTION

The temporal factors of speech have a major role in speech understanding, speech meaning, and is of a great importance for NLP (natural language processing) and for speech technology applications, such as TTS (text-to-speech synthesis), ASR (automatic speech recognition) widely used in infocommunication. Research on vowel duration has primarily been focused on the issues of whether and how various factors affect the timing of segments (in various languages). In determining the physical duration of a speech sound, the identity (including the spectral properties) of the segment plays a decisive role, but so do stress, fundamental frequency, phonetic position, syllable structure, speech sound context, full length of the utterance, syntactic boundaries, the speaker’s physiological and psychological state, word frequency, etc. (see [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]). In Hungarian, there are five pairs of vowels [u – w; o – ó, y – γ; ő – Ś, i – i], the members of which hardly differ in quality or not at all. Short and long members of those five pairs are contrastive, i.e., they exhibit phonological opposition (e.g., őrül ‘s/he is glad’ – őrül ‘s/he is going crazy, bor ‘wine’ – bór ‘boron’, irat ‘document’ – írat ‘write-causeative’). Phonologically long vowels of a language are not necessarily realized longer than phonologically short ones in actual pronunciation, especially in spontaneous speech. The reason is that the objective duration of vowels is determined by a number of diverse factors that often interact to yield the phonetic duration of the given sound in the given context. Phonetic realizations of phonological length are accordingly rather varied, with objective temporal values typically exhibiting some overlap between categories. Such overlapping durations mean that a phonologically short vowel may occasionally be realized with a longer duration than a long one (and vice versa). Word stress, in Hungarian, invariably falls on the initial syllable, irrespective of whether the first syllable of the word contains a short or a long vowel [12, 13]. Due to the agglutinative character of the language, Hungarian words consist of 3.5 syllables on average in spontaneous speech. Temporal relationships of Hungarian vowels have been analyzed in the past few decades from various perspectives (e.g., [4, 14, 15, 16, 17]). In Hungarian, WER (word error rate) of recognition of continuous speech is about 50–60% [18]. Tóth and Kocsor [19] showed that explicit duration modeling decreased WER in read speech based on an HMM/ANN system (hidden Markov-model/artificial neural network). However, no automatic prediction of vowel durations has been attempted in spontaneous Hungarian speech so far.

In general, automatic prediction of segment duration can be solved either by rule-based or statistics-based methods. Rule-based methods start from intrinsic vowel duration and vowel durations are modified by the applied rules [20, 21]. Klatt’s work [21] presents a rule-based method for predicting vowel duration, which can be described by the following equation (1):

\[
\text{Dur} = \frac{(\text{INDUR} - \text{MINDUR}) \times \text{PRCNT}}{100} + \text{MINDUR}
\]

where INDUR is the inherent duration, MINDUR is a certain minimal duration, PRCNT is the value of context dependent parameter. Rule-based models often seem to be
overgeneralized, and exceptions can be handled only by rather complex solutions [22].

In the case of a small corpus rule-based systems offer high precision, but in the case of a large corpus these algorithms require a lot of time-consuming manual work. Moreover, results are not suitable for generalization. Another approach is the class of statistics-based methods to predict the vowel duration. The common disadvantage of these methods is that to learn the rules it is necessary to have a large phonetically labeled database. Statistics-based methods have two sub-classes: parametric and nonparametric regression models [23]. Parametric regression models include the sums-of-products model (SOP) [24, 25], the generalized linear model (GLM) and the multiplicative model (MM) [26]. The nonlinear regression models can be either classification-and-regression-tree models (CHART) or artificial neural network models (ANN) [22, 26, 27, 28].

The sums-of-products model is a popular duration modeling method [24, 25] whose construction consists of the following three steps. (i) Building a category tree, i.e. to construct the category tree we need a priori knowledge about duration ratios. We assume that the factors and parameters have a same effect on each segment which is represented by one leaf. (ii) The second step is creating a SOP model for each category tree leaf. (iii) The third step is predicting the parameter estimation. This algorithm has a good predicting capability; it means that the correlation between the predicted and the measured duration is high. The disadvantage of this method is that to create the category tree, precise language knowledge about the effect of factors on the segment duration is needed [27].

The most popular algorithm predicting segment duration is the nonlinear regression machine learning method known as the classification and regression tree [29], which is a statistical modeling technique. CART can produce a relationship between dependent variable y corresponding to feature vector X [20, 24, 25, 30]. The disadvantage of this method is that it does not provide sufficiently accurate results. However, this method has many advantages, for example it can measure the effect of the factors on the feature vectors [27].

The other machine learning method used to predict segment duration is the artificial neural network. The ANN (artificial neural network) is well applicable in fitting functions. In fact, there is proof that a fairly simple neural network can fit any practical function. The ANN method is most commonly used to model syllable duration as well [31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. Generally, a two-layer feedforward network, utilizing a sigmoid transfer function in the hidden layers and a linear transfer function in the output layers is used in fitting the data.

In the present paper, we investigate regression trees and neural networks used to automatically predict vowel durations based on vowel quality and vowel formants, phonological context and speaker identity and spectral features as well.

A. Speech database

The Hungarian phonetically-based multi-purpose database, identified by the acronym BEA (BÉszélőnyelvi Adatbázis ‘spoken language database’) has accumulated a large amount of spontaneous speech of various types (narratives, story recalls, presentation of opinion on a given topic and conversations). In addition, for purposes of comparison, it contains repetitions of sentences of various lengths and texts read aloud by each speaker. The spontaneous speech samples are on average 40 minutes long per person. The structured material of BEA presently amounts to 260 hours produced by 280 adult Budapest speakers (aged between 20 and 85, 152 females and 128 males), providing annotated material for various types of research and practical applications. Each subject’s speech was recorded in the same sound-attenuated room using a unidirectional high-quality microphone and a digital recorder connected to a computer. The recording environment and the technical facilities were the same in all cases. (For further information see http://www.nyutud.hu/adath/bea/index.html).

B. Method

Spontaneous speech by 10 monolingual Hungarian-speaking young subjects (5 men and 5 women, aged between 25 and 34) was selected from the BEA database. All speakers had normal hearing and no known speech defect. We analyzed a total of over 110 minutes of speech samples (8,214 words). The speech rate of the subjects (including silent and filled pauses) was roughly uniform; they spoke 126 words per minute, on average. We analyzed a total of 5,837 vowels. The speech material was first annotated at the phrase level using the Praat software [41]; then an automatic segmenter was used (ftp://ftp.bas.uni-muenchen.de/pub/BAS/SOF TW/MAUS) to annotate it at the segment level. The MAUS software delimits realizations of the individual phonemes using the sound delimitation criteria generally employed in phonetic analysis (Fig. 1). (The software had been originally trained on German; we used it with visual and auditory feedback.) Automatic annotation was manually corrected where necessary by the two authors separately. The segment boundaries were identified using waveform and spectrum, and the onset and end of vowel formants were considered as cues.

\[\text{Figure 1. Annotated spectrogram of 'ez nagyon fontos' 'it is so important' (Q marks the vowel of the initial syllable, X marks that of the final syllable).}\]
II. PREDICTION OF VOWEL DURATIONS BASED ON CART

A. Classification and Regression Tree

The function at the heart of the CART method is the regression function which creates a relationship between the variables by a closed mathematical function; that means that if the feature vector is known the dependent variable can be predicted. It is a kind of supervised learning. In the case of the training phase the task is to find a function \( d(x) \) that describes the mathematical relationship between independent variable data elements in \( X \) and dependent variable \( Y \). The training sets \( L \) contain \( n \) observations: \( L = \{ (x_1, y_1), \ldots, (x_n, y_n) \} \) [20]. The means squared error is used to find \( d(x) \) which can be expressed by the following equation: \( E(\hat{d}(x) - \hat{y}(x))^2 \) for regression, where \( \hat{y}(x) \) is the expected value of \( y \) at \( x \). This regression function is not only used in the regression tree but also in the classification tree which a constant or a regression function is fitted to the data [20].

Our starting point in building the regression tree contains only a root node \( t_1 \) that includes all data elements in the training set \( L \). The problem is to detect the optimal binary split of \( t_1 \) into \( t_{L_1}, t_{R_1} \). For a measured valued feature \( i \) all splits of the form \( x_i < \tau \) \( (\tau \in \mathbb{R}) \) would be tested. Once an optimum split of \( t_1 \) is found, the same procedure is applied to the two child nodes \( t_{L_1}, t_{R_1} \). This recursive algorithm uses a stopping criterion like for example the size of a node is less than some specified minimal size or the decrease of prediction error is observed [20].

For each split \( \tau \) of the form \( x_i < \tau \) \( (\tau \in \mathbb{R}) \) the error estimate of node \( t \) [20] would be tested. Once \( t \) \( (\tau \in \mathbb{R}) \) is found, the same procedure is applied to the two child nodes \( t_{L_1}, t_{R_1} \). This recursive algorithm uses a stopping criterion like for example the size of a node is less than some specified minimal threshold or the decrease of prediction error is observed [20]. The optimal split and prognosis inside terminal nodes depend on \( R(T) \), the error estimate of the tree \( T \) and \( R(t) \), the error estimate of node \( t \) [20]. Two measurements were applied in this article: root mean square error (RMSE) and mean relative error (MRE).

\[
R_{RMSE}(T) = \sqrt{\frac{1}{N} \sum_{i=1}^{M} R_{RSE}(t_i)}
\]

where \( RRE \) is

\[
R_{RSE}(t) = \sum_{x \epsilon t} (y_n - y(t))^2
\]

Mean relative error (MRE):

\[
R_{MRE}(T) = \frac{1}{N} \sum_{i=1}^{M} R_{RE}(t_i)
\]

where \( R_{RE} \) is

\[
R_{RE}(t) = \sum_{x \epsilon t} \left| y_n - \hat{y}(x) \right| y_n
\]

Let \( S \) denote the set of all probable splits of a node \( t \). For each split \( s \epsilon S \) of \( t \) into \( t_{L_1}, t_{R_1} \) let

\[
\Delta R(s, t) = R(t_L) - R(t_R)
\]

Then the best split \( s^* \epsilon S \) would be

\[
s^* = \text{argmax}_{s \epsilon S} \Delta R(s, t)
\]

The regression tree is built by recursively splitting nodes. In other words, in this processing the algorithm maximizes the reduction in \( R(T) \). The initial value (RMSE) of \( y(t) \) which minimizes \( R_{RSE}(t) \) is the mean of all values \( y_n \) falling into node \( t \) [20]:

\[
y(t) = \bar{y}(t) = \frac{1}{N(t)} \sum_{x \epsilon t} y_n
\]

The other measurement is the MRE. In this algorithm the value of \( y(t) \), which minimizes \( R_{RE}(T) \) is the median of all values \( y_n \) falling into node \( t \) [20].

B. Features used to predict vowel duration

There are characteristic features used at the segment and syllable levels. Those at the segment level are target vowel identity; phonological class of the preceding and the following segment (example: voiced/voiceless); phonological vowel length; speaker identity; those at the syllable level are vowel position in the word (first syllable, medial position, final position).

C. Experimental setup of CART

In this work we use a 10-fold cross-validation processing both for training and testing. In this case the corpus contains a training set using 90 % of the database, and 5-5% of the database for both validation and testing. Various models were constructed for data elements, and all features were trained to predict vowel duration.

D. Results

The results show that the RMSE value is 33 ms when a classification and a regression tree are used for predicting vowel duration. The correlation value between measured and predicted data is 0.57. The median of the measured data is 70.5 ms, and the standard deviation is 37.69 ms. The median of the predicted data is 71.86 ms, and the standard deviation is 29.98 ms.

III. PREDICTING VOWEL DURATIONS BASED ON FFNN

A. Feedforward Neural Network

A two-layer feedforward neural network (FFNN) is applied for predicting the durations of vowels. In the training phase the FFNN finds the relationship between the input and the target feature vectors of the data elements. In the present study, the size of the input vector is 29 and the size of the target vector is 1.
It is well known that any continuous vector-valued function can be realized by an ANN that contains two hidden layers [42]. The present network for data fitting is a feedforward network with a tan-sigmoid transfer function in the hidden layer and a linear transfer function in the output layer (Fig. 3). The segment positional and contextual factors are the input vector of the ANN. Before the training phase features are usually normalized, therefore feature vector elements were transformed linearly into the range [-1, +1]. The backpropagation method was applied to tune the weights of the FFNN in the training cycle to reduce the MSE for all vowel durations [43, 44].

B. Features used to predict vowel duration

There are characteristic features used at both the segment and syllable levels. Those used at the phone level are target vowel identity; phonological class of the preceding and the following phones (example: voiced/voiceless); phonological vowel length; speaker identity; those used at the syllable level are vowel position in the word (first syllable, medial position, final position).

C. Results

The best result was obtained by FFNN with 20 hidden layers. The iteration stopped at the epoch 5 where the validation performance is 0.016 (Fig. 4).

The results show that the RMSE value was 29 ms when a FFNN was used to model vowel duration. The correlation value between the measured and predicted data was 0.62 (Fig. 5). The median of the measured data was 70.5 ms, and the standard deviation was 37.69 ms. The median of the predicted data was 71.86 ms, and the standard deviation was 29.98 ms.

IV. MODELING VOWEL DURATIONS BASED ON FFNN AND SPECTRAL FEATURES

A. Spectral features used to predict vowel duration

We measured the first three formant values of the vowel and the right-left neighboring sounds. The LPC coefficient with the algorithm by Burg (time step: 5 ms; Hamming window length: 25 ms; maximum frequency: 5500 Hz; pre-emphasis frequency: 50 Hz) was used in formant measurements. On the left and right sounds the CoG (Center of Gravity) and the energy difference between 0–500 Hz and 500–4000 Hz were measured. The spectral feature analyzed was measured by LPC analysis (16 prediction order, Hamming window length: 25 ms; time step: 5 ms; pre-emphasis frequency: 50 Hz).

B. Experimental setup of FFNN

To predict vowel duration with the help of characteristics and spectral features we used FFNN algorithm. The feature vector length was 36, therefore the number of hidden layer in FFNN was 36 layers. To prepare the spectral data we normalized them by the standardized z-score algorithm.

C. Results

The best result was obtained by FFNN with 36 hidden layers. The iteration stopped at epoch 12.

The results show that the RMSE value is 25 ms when a feedforward neural network is used to predict vowel duration with characteristic features combined with the spectral features. The correlation value between the measured and predicted data is 0.78 (Fig. 6). The median of the measured data is 70.5 ms, and the standard deviation is 37.69 ms. The median of the predicted data is 71.12 ms, and the standard deviation is 36.56 ms.
V. COMPARISON OF THE RESULTS OBTAINED BY THE THREE MODELS

The average prediction error ($\mu$), standard deviation ($\sigma$) and regression coefficient ($r$) were measured to compare the result of the different models. Computation of the average prediction error is shown by the following equation [40]:

$$\mu = \frac{\sum|x_i - \bar{x}|}{N}$$  \hspace{1cm} (9)

and the standard deviation:

$$\sigma = \sqrt{\frac{\sum d^2}{N}}, \text{ and } d^2 = (x_i - y_i) - \mu$$  \hspace{1cm} (10)

where $x_i$ is the measured and vowel duration, $y_i$ is the predicted vowel duration, and $N$ is the number of vowels.

The results show that the FFNN based on characteristic and spectral features offered the best prediction. The worst result was obtained by the CART method (Fig. 1).

A statistical analysis using ANOVA was prepared to investigate the deviation of predicted vowel durations from the corresponding measured value ($x_i - y_i$), comparing the three methods. The results show that there is a significant difference between the prediction capability of the tree methods. ($F(2, 17511) = 193,330; p<0.001^{**}$; Tukey post hoc tests showed also significant differences.)

VI. SUMMARY AND CONCLUSION

Human speech production is one of the most complicated cognitive human processes. Therefore, in modeling speech production different fields of sciences must cooperate. One of the major goals of infocommunication is to build connections between cognitive science and various engineering technologies. In this research too, the authors combined the area of cognitive science of linguistics (represented by phonetics and phonology) and engineering (represented by speech technology).

The aim of the present paper was to investigate different supervised learning methods predicting vowel duration, based on different feature vectors consisting of characteristic and spectral features. Positional, contextual, and phonological attributes of the vowel constituted vowel characteristic features. Spectral features, such as the first three formants of the vowel and the CoG of the left-right neighboring sounds and spectral energy difference data were measured and included in the feature vector.

Classification and Regression Tree and Feedforward Neural Network supervised learning method were applied for predicting the vowel duration. We built three different models to predict the vowel duration. (i) CART was trained and tested by characteristic features, (ii) FFNN was trained and tested by characteristic features; (iii) FFNN was trained and tested by characteristic features combined with spectral features. The results of the different learning methods were compared by calculating the average prediction error ($\mu$), standard deviation ($\sigma$) and correlation coefficient ($r$) between the predicted and actual durations of the vowels. The best result was obtained by the FFNN using combined features. This result indicated that the prediction error can be further minimized by combining characteristic and spectral features. However, a minor increase of accuracy of FFNN methods over CART comes from the fact that neural networks can weight the feature vector according to the output.

For further research, the authors suggest that performance can be further improved using other features, such as speaker identity, speaker age and prosodic features as well. Other factors affecting the prediction rate may be preciseness of annotation, heterogeneity of data, and topology of the FFNN, etc.

Our method to predict the vowel duration can be applied primarily in text-to-speech synthesis, but also can be used in speech recognition, language identification, natural language processing.

ACKNOWLEDGMENT

We would like to thank D. Gardner-Bonneau and F. Nolan for their valuable comments on an earlier version of this paper.

REFERENCES


![Figure 6. Correlation between measured and predicted data in the case of training tests, validation tests and all of them.](image-url)


