

Automatic classification of segmental second language speech quality using prosodic features

Eero Väyrynen¹, Heikki Keränen², Juhani Toivanen³ and Tapio Seppänen⁴

^{1,2,4}MediaTeam, University of Oulu

³MediaTeam, University of Oulu & Academy of Finland

Abstract

An experiment is reported exploring whether the general auditorily assessed segmental quality of second language speech can be evaluated with automatic methods, based on a number of prosodic features of the speech data. The results suggest that prosodic features can predict the occurrence of a number of segmental problems in non-native speech.

Introduction

Our research question is: is it possible, by looking into the supra-segmentals of a second language variety, to gain essential information about the segmental aspects, at least in a probabilistic manner? That is, if we know what kinds of supra-segmental features occur in a second language speech variety, can we predict what some of the segmental problems will be?

The aim of this research is to find if supra-segmental speech features can be used to construct a segmental model of Finnish second language speech quality. Multiple nonlinear polynomial regression methods (for general reference see e.g. Khuri (2003)) are used in an attempt to construct a model capable of predicting segmental speech errors based solely on global prosodic features that can be automatically derived from speech recordings.

Speech data

The speech data used in this study was produced by 10 native Finnish speakers (5 male and 5 female), and 5 native English speakers (2 male and 3 female). Each of them read two texts: first, a part of the Rainbow passage, and second, a conversation between two people. Each rendition was then split roughly from the middle into two smaller parts to form a total of 60 speech samples (4 for each person). The data was collected by Emma Österlund, M.A.

Segmental analysis

The human rating of the speech material was done by a linguist who was familiar with the

types of problems usually encountered by Finns when learning and speaking English. The rating was not based on a scale rating of the overall fluency or a part thereof, but instead on counting the number of errors in individual segmental or prosodic units. As a guideline for the analysis, the classification by Morris-Wilson (1992) was used to make sure that especially the most common errors encountered by Finns learning English were taken into account.

The main problems for the speakers were, as was expected for native Finnish speakers, problems with voicing (often with the sibilants), missing friction (mostly /v, θ, ð/), voice onset time and aspiration (the plosives /p, t, k, b, d, g/), and affricates (post-alveolar instead of palato-alveolar). There were also clear problems with coarticulation, assimilation, linking, rhythm and the strong/weak form distinction, all of which caused unnatural pauses within word groups.

The errors were divided into two rough categories, segmental and prosodic, the latter comprising any unnatural pauses and word-level errors – problems with intonation were ignored. Subsequently, only the data on the segmental errors was used for the acoustic analysis.

Acoustic analysis

For the speech data, features were calculated using the f0Tool software (Seppänen et al. 2003). The f0Tool is a software package for automatic prosodic analysis of large quanta of speech data. The analysis algorithm first distinguishes between the voiced and voiceless parts of the speech signal using a cepstrum based voicing detection logic (Ahmadi & Spanias 1999) and then determines the f0 contour for the voiced parts of the signal with a high precision time domain pitch detection algorithm (Titze & Haixiang 1993). From the speech signal, over forty acoustic/prosodic parameters were computed automatically. The parameters were:

- A) general f0 features: mean, 1%, 5%, 50%, 95%, and 99% values of f0 (Hz), 1%- 99% and 5%-95% f0 ranges (Hz)
- B) features describing the dynamics of f0 variation: average continuous f0 rise and fall (Hz), average f0 rise and fall steepness (Hz/cycle), max continuous f0 rise and fall (Hz), max steepness of f0 rise and fall (Hz/cycle)
- C) additional f0 features: normalised segment f0 distribution width variation, f0 variance, trend corrected mean proportional random f0 perturbation (jitter)
- D) general intensity features: mean, median, min, and max RMS intensities, 5% and 95% values of RMS intensity, min-max and 5%-95% RMS intensity ranges
- E) additional intensity features: normalised segment intensity distribution width variation, RMS intensity variance, mean proportional random intensity perturbation (shimmer)
- F) durational features: average lengths of voiced segments, unvoiced segments shorter than 300ms, silence segments shorter than 250ms, unvoiced segments longer than 300ms, and silence segments longer than 250ms, max lengths of voiced, unvoiced, and silence segments
- G) distribution and ratio features: percentages of unvoiced segments shorter than 50ms, between 50-250ms, and between 250-700ms, ratio of speech to long unvoiced segments (speech = voiced + unvoiced<300ms), ratio of voiced to unvoiced segments, ratio of silence to speech (speech = voiced + unvoiced<300ms)
- H) spectral features: proportions of low frequency energy under 500 Hz and under 1000 Hz

dre polynomials and cross terms to introduce nonlinearity:

$$P_{1p} = x_p,$$

$$P_{2p} = \frac{1}{2}(3x_p^2 - 1),$$

$$P_{3pq} = x_p x_q, \quad q < p.$$

The resulting total of 1127 new features P was then searched to find the best performing regression coefficients to be described next.

A sequential forwards-backwards floating search (SFFS) (Pudil et al. 1994) was used to find a set of 15 best features $a_k \in P$ by minimising the sum of squared errors (SSE) given by solving a standard multiple linear regression procedure:

$$y_i = \beta_0 + \beta_1 a_{i1} + \beta_2 a_{i2} + \dots + \beta_k a_{ik} + \varepsilon_i,$$

where $i = 1, 2, \dots, n$ are independent samples, β_k the regression parameters, and ε_i is a random error term. An LMS solution

$$(\mathbf{A}^T \mathbf{A}) \hat{\boldsymbol{\beta}} = \mathbf{A}^T \mathbf{y}$$

was used to produce regression estimates

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 a_{i1} + \hat{\beta}_2 a_{i2} + \dots + \hat{\beta}_k a_{ik}.$$

The best features were then transformed using a robust PCA method (Hubert et al. 2005) to remove any linear correlations. The PCA transformed 15 features (scaled to $-1 \leq x \leq 1$ interval) were then searched again with SFFS to select a set of 8 final PCA transformed features.

The motivation for this process was to combat any over learning of data by limiting the number of resulting regression coefficients to as small a number as possible. A person independent hold out cross-validation process was also used throughout the feature selection procedure to ensure generalization of the resulting models. In the hold out procedure for each person his or her samples were rotated out from the database in turn and a regression model was trained using the remaining people's samples. The resulting models were then used to predict the corresponding set of samples held out.

The final regression model was then validated by inspecting the total cross-validated regression residual and each of the individual p values of the final cross-validation training round. The p values represent the probability

Nonlinear regression

A multiple polynomial nonlinear regression of segmental speech errors was performed. First, the x_p , $p = 1, 2, \dots, 46$ raw prosodic features were scaled to 0-mean and scaled to $-1 \leq x \leq 1$ interval. The scaled features were then transformed using the first and second order legen-

the data is drawn from a distribution consistent with the null hypothesis where the prosodic data is assumed containing no explanatory linear components at all. Finally, the consistency of the speaker independent regression coefficients was inspected to ensure the validity and stability of the model.

Results

The first feature selection resulted in a feature vector that contains no second order polynomials. The 15 features are described in Table 1. A selected cross term is indicated as “*feature X feature*”.

Table 1. Selected features in the first search

trend corrected mean proportional random f0 perturbation (jitter)
average lengths of unvoiced segments longer than 300ms
normalised segment intensity distribution width variation
50% values of f0 X percentages of unvoiced segments between 250-700ms
1% values of f0 X max lengths of silence segments
average continuous f0 rise X average continuous f0 fall
average continuous f0 fall X max steepness of f0 fall
average continuous f0 fall X max lengths of voiced segments
max continuous f0 fall X ratio of speech to long unvoiced segments (speech = voiced + unvoiced<300ms)
max steepness of f0 rise X ratio of speech to long unvoiced segments (speech = voiced + unvoiced<300ms)
RMS intensity variance X average lengths of silence segments longer than 250ms
unvoiced segments longer than 300ms X percentages of unvoiced segments shorter than

50ms

max lengths of voiced segments X max lengths of silence segments

max lengths of voiced segments X normalised segment f0 distribution width variation

ratio of silence to speech (speech = voiced + unvoiced<300ms) X normalised segment f0 distribution width variation

No second order legendre polynomials were included by the selection procedure while many cross terms were included in the regression. The omission of second order nonlinearity and heavy reliance of cross terms suggests that differentiating information is perhaps coded more as co-occurrences of features rather than strict nonlinear combinations of raw prosodic features.

After the robust PCA transformation a feature vector containing 8 best linearly independent features was selected as the final regression model. The resulting regression coefficients of each person independent training round were found to be consistent with little or no variation. The final models relevant person independent cross-validation training regression statistics are shown in Table 2. In the table, a range of R^2 and p value for the training with cross-validated R^2 and p value for the testing are shown. The p values indicate that the null hypothesis can be rejected i.e. prosodic features do contain a model capable of predicting speech proficiency. The R^2 values further shows that a majority of the variance is explainable by the chosen features.

Table 2. Final regression statistics

Person	R^2	p
training	0.752 – 0.858	<0.001
cross-validated	0.659	<0.001

The cross-validated residual of the final 8 robust PCA feature regression is shown in Figure 1 and the corresponding scatterplot of human and regression estimate of errors in Figure 2.

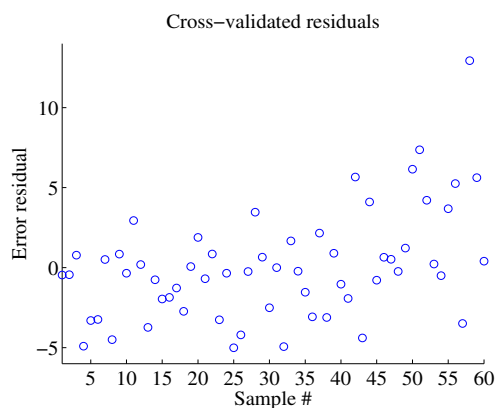


Figure 1. Cross-validated regression residuals. The data is ordered in an ascending human error with the circles indicating the residual errors.

The resulting regression residuals indicate that perhaps some nonrandom trend is still present. Some relevant information not included by the regression model is therefore possibly still present in the residual errors. It may be that the used prosodic features do not include this information or that more coefficients in the model could be justified.

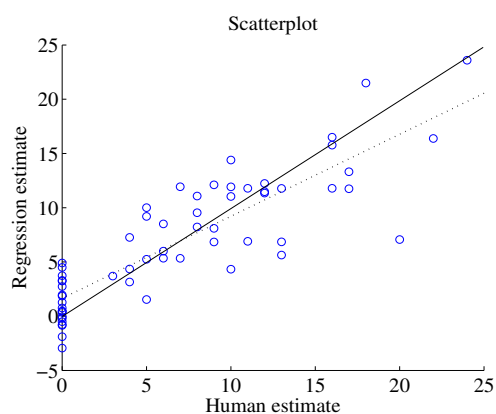


Figure 2. A scatterplot of human errors against the corresponding cross-validated regression estimates. The solid line shows for reference where a perfect linear correspondence is located and the dashed line is a least squares fit of the data.

The scatter plot shows a linear dependence of 0.76 between human and regression estimates with 66% of variance explained.

Conclusion

The results suggest that segmental fluency or “correctness” in second language speech can be modelled using prosodic features only. It seems that segmental and supra-segmental second language speech skills are interrelated. Parameters describing the dynamics of prosody (notably, the steepness and magnitude of f_0 movements – see Table 1) are strongly correlated with the

evaluated segmental quality of the second language speech data. Generally, it may be the case that segmental and supra-segmental (prosodic, intonational) problems in second language speech occur together: a command of one pronunciation aspect may improve the other. Some investigators actually argue that good intonation and rhythm in a second language will, almost automatically, lead to good segmental features (Pennington, 1989). From a technological viewpoint, it can be concluded that a model capable of estimating segmental errors can be constructed using prosodic features. Further research is required to evaluate if a robust test and index of speech proficiency can be constructed. Such an objective measure can be seen as a speech technology application of great interest.

References

- Ahmadi, S. & Spanias, A.S. (1999) Cepstrum based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transaction on Speech and Audio Processing* 7 (3), 333–338.
- Hubert, M., Rousseeuw, P.J., Vanden Branden, K. (2005) ROBPCA: a new approach to robust principal component analysis. *Technometrics* 47, 64–79.
- Khuri, A.I. (2003) *Advanced Calculus with Applications in Statistics*, Second Edition. Wiley, Inc., New York, NY.
- Morris-Wilson, I. (1992) *English segmental phonetics for Finns*. Loimaa: Finn Lectura.
- Pennington, M.C. (1989) Teaching pronunciation from the top down. *RELC Journal*, 20–38.
- Pudil, P., Novovičová, J. & Kittler J. (1994) Floating search methods in feature selection. *Pattern Recognition Letters* 15 (11), 1119–1125.
- Seppänen, T., Väyrynen, E. & Toivanen, J. (2003). Prosody-based classification of emotions in spoken Finnish. *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH-2003* (Geneva, Switzerland), 717–720.
- Titze, I.R. & Haixiang, L. (1993) Comparison of f_0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research* 36, 1120–1133.