Laboratório de Fonologia
Universidade Federal de Minas Gerais
Faculdade de Letras • sala 3104

CEFALA

POSLIN·UFMG FALE FACULDADE DE LETRAS UFMG

Articulatory and vocal speaker variability in connected speech

Maria Mendes Cantoni (Universidade Federal de Minas Gerais, mmcantoni@gmail.com)
Adelino Pinheiro Silva (Police Academy of Minas Gerais, adelinocpp@gmail.com)

1 INTRODUCTION

Sources of variation in speech [1,2]

- Linguistic variation: phonetic-phonological, coarticulatory
- Speaker-related variation: sociolinguistic, personal

Between speaker variability: anatomical or physiological (difference on vocal tracts or difference on motor routines used by different speakers)

Within-speaker variability: biomechanical (differences on how speech movements are actually implemented by the same individual.)

Problems

1) The role of different components of the vocal tract in speaker identification is not clear [3].
2) Only a few studies on speaker variability have used connected speech [4].

In this study

Aiming to model speaker-related variability, we consider the roles of articulatory structures and voice in speaker classification in connected speech. Questions addressed:

- How much speaker variation is due to articulation differences and how much is due to voice differences?
- Which acoustic measures are more robust for speaker identification in connected speech?

We designed a regression based classification procedure, in which linguistic predictable variability is removed and the residuals are used as the input for variability modeling [cf. EQ1]. Besides allowing us to work with a small data set and spontaneous speech, this procedure would still preserve the original variables of interest.

$$X_n \approx V_{C|n} + (V_{S|n} + V_{NC|n} + \varepsilon_n) \rightarrow X_n \approx V_{C|n} + \varepsilon_{S,NC|n} \quad \text{(EQ 1)}$$

X_n = random variable for acoustical measure n; $V_{S|n}$ = speaker variability for n; $V_{C|}$ = context variability for n; $V_{C|}$ = variability for n not accounted by the context; ε_n = residuals; $\varepsilon_{S,NC|n}$ = residuals including speaker variability and unaccounted variability.

2 METHODS

Materials

- Vowels in spontaneous speech
- Recordings from 18 speakers from CEFALA-1, a Brazilian Portuguese data set [5]
- Manual segmentation and labelling in Praat [6]

Data Coding

- Linguistic variables relevant to the language sound system: preceding and following context, vowel quality, nasality, stress degree, number of syllables, syllable structure

Measurements

- Acoustic measurements after [3], divided into articulatory and vocal and estimated in mean and variation coefficient: duration, formants (F1, F2, F3, F4 and dispersion), intensity, f0, spectral slope at four spectral regions, SNR, CPP

Modelling **[cf. FIG 1]**

A - Selection of vowels common to all speakers

B - Feature extraction (measurements)

C - Bootstrapping

D - Data split into training (70%) and test sets

E - Generalized linear regression model (GLM) fit, with linguistic variables as factors

Using the normalized Euclidean-distance of residuals from GLM (GLM_RES):

F - Logistic regression model fit (training set only)

G - Logistic inference for same (SS) and different speaker (DS) classification

Model evaluation

- Empirical cross-entropy (ECE, [7]) calculation for evaluation of model fit
- Multi-dimensional scaling (MDS) for comparison with raw data and also PCA from raw data

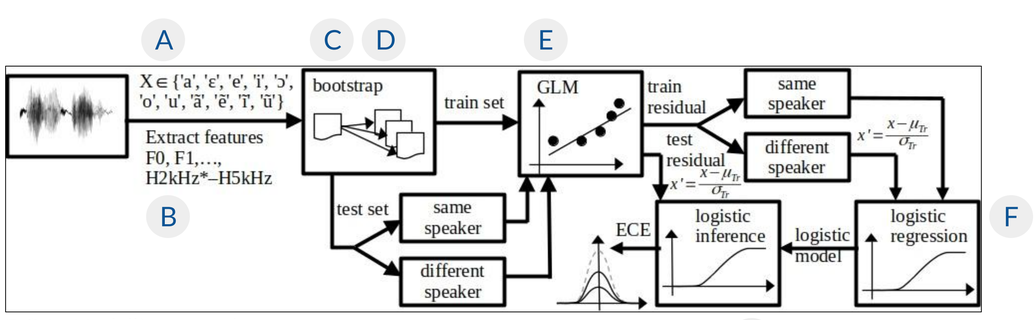

FIG 1 - Modelling procedures

3 RESULTS

Most between-speaker acoustic variability comes from sex-related differences in vocal tract

[FIG 2] PCA from raw data:

- Clear separation between female (F) and male (M) voices
- 9 PCs account for 95% variance

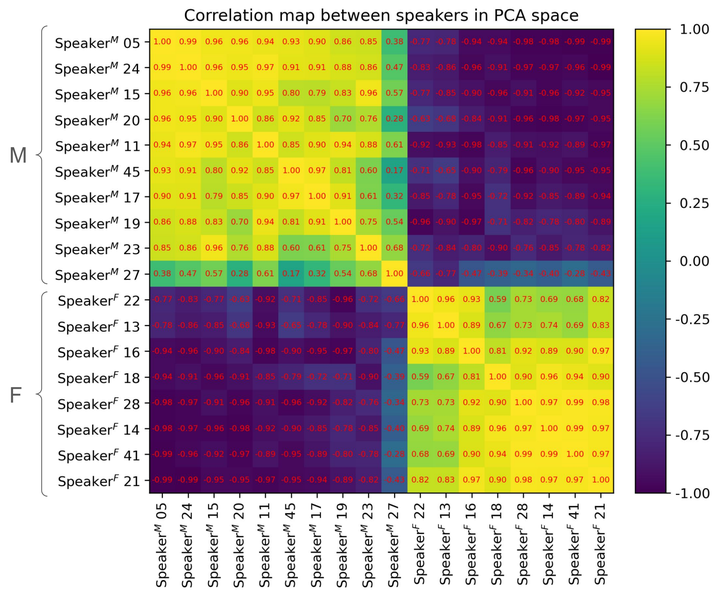

FIG 2 - Speakers correlation map using PCA

Most acoustic variables alone are able to separate at least one speaker from the group

[FIG 3, left] Anova by speaker on raw data:

- All but 3 variables have p > 0.05 (dash-dot line)

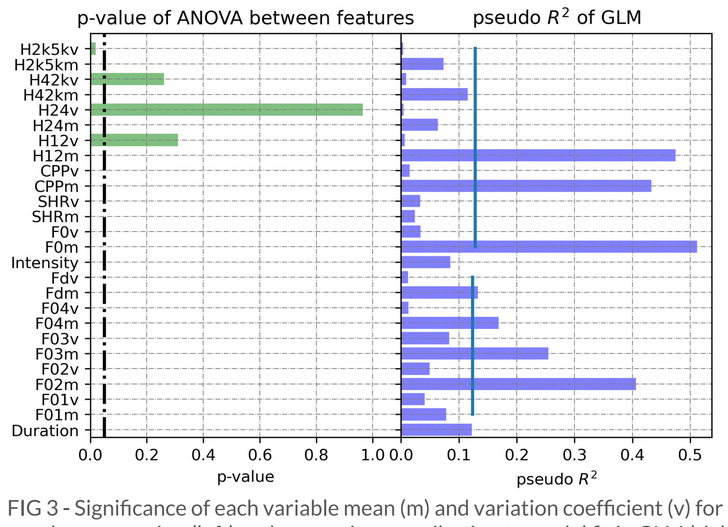

FIG 3 - Significance of each variable mean (m) and variation coefficient (v) for speaker separation (left) and respective contribution to model fit in GLM (right).

Articulatory and vocal variables are relevant to speaker classification

[FIG 3, right] Variables with best model fit on GLM:

- 7 variables have R^2 > mean R^2 (blue vertical line):
  - 3 vocal variables: variation of H*1-H*2 and CPP and mean F0
  - 4 articulatory variables : mean formant dispersion, mean F4, mean F3 and mean F2
- Articulatory variables with mean R^2 = 0.124 and vocal variables with mean R^2 = 0.129

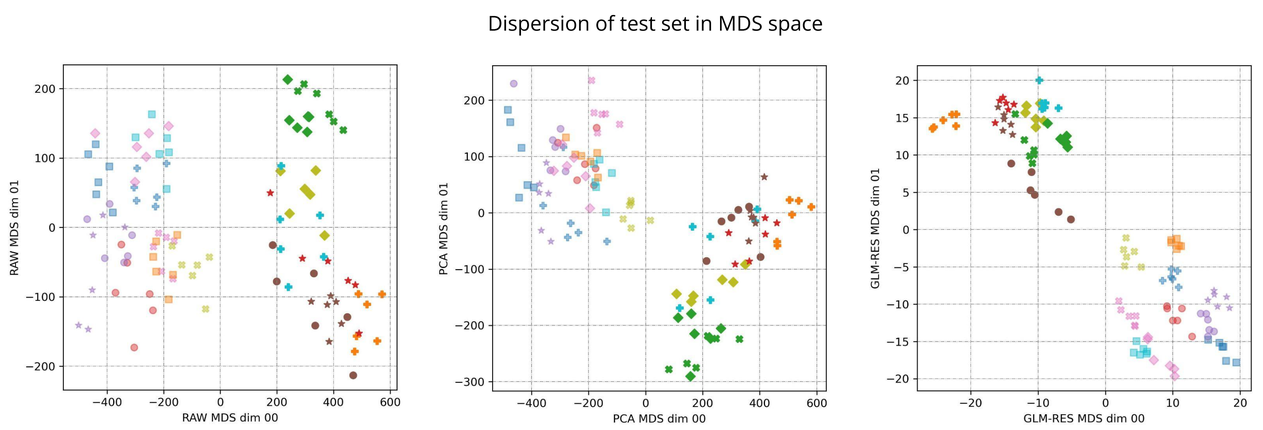

FIG 4 - Multi-dimensional scaling comparative analysis of classification. Left to right: on raw data, on PCA on raw data and on the residual space from GLM. Each speaker is displayed with a singular shape-color of dots. Female speakers shown in solid dots and male speakers in transparent dots.

The regression model designed has a good performance at speaker classification

[FIG 4] Comparison of classification based on GLM-RES with a classification based on raw data and PCA

- Speaker classification based on raw (RAW) data and on PCA separates female and male speakers, but points from different speakers are overlapped
- The regression method designed (GLM-RES) is able to fully separate individual speakers, with minimum overlap

Both articulatory and vocal variables, and articulatory outperforms vocal

[FIG 5] Comparison of classification based on GLM-RES with a classification based on raw data and PCA

- The regression model (GLM-RES) performed better (with EER = 4.1%) than PCA and raw data (RAW)
- The model containing both articulatory and vocal variables (GML-RES) performed better than articulatory-only (-ART) and vocal-only similar models (-VOC), in this order

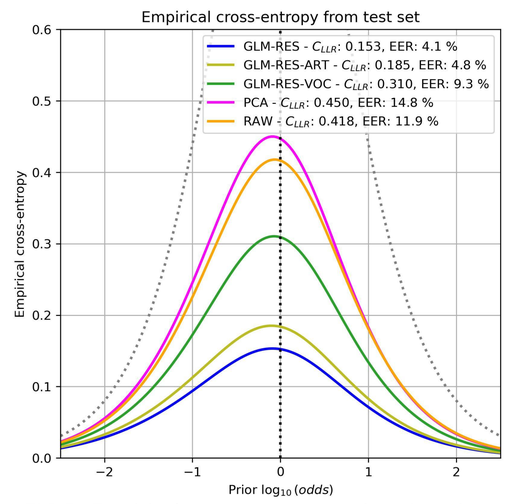

FIG 5 - Empirical cross-entropy as a function of the prior log odds of different classification procedures: based on the raw data (orange), PCA (pink), the regression model with articulatory variables only (khaki) and vocal variables only (green) and the full regression model (blue), compared to the classification at chance (dotted line). CLLR (log-likelihood ratio cost) and EER (equal error rate) provided for each procedure.

4 FINAL REMARKS

- Limitations: dependent only static variability addressed, not dynamic. Does not account for consonant variability.
- Future work: Improve the procedure to model dialect variability. Expand data set size. Implement automatic segmentation, alignment and annotation

SCAN
For poster and complete reference list

REFERENCES
[1] Ladefoged and Broadbent (1957)
[2] Kilbourn-Ceron and Goldrick (2021)
[3] Lee, Keating and Kreiman (2019)
[4] Lee and Kreiman (2022)
[5] Yehia, Follador and Silva (2019)
[6] Boersma and Weenink (2023)
[7] Ramos and Gonzalez-Rodriguez (2013)

Acknowledgments: We are grateful to Prof. Jody Kreiman and Prof. Hani Yehia, for comments on previous versions of this work. All the remaining problems are our own.