University of Debrecen, Institute of Psychology, Hungary
Budapest University of Technology and Economics (BME TMIT), Hungary

From text to formants – indirect model for trajectory prediction based on a multi-speaker parallel speech database

Kálmán Abari, Tamás Gábor Csapó, Bálint Pál Tóth, Gábor Olaszy
abari.kalman@arts.unideb.hu, {csapot, toth.b, olaszy}@tmit.bme.hu

1. Introduction

An indirect HMM-based model is presented capable of estimating formant trajectories from Hungarian text (Text-to-Formant conversion, TTF).

- **Model input:** a Hungarian sentence (text)
- **Model output**: phonetically correct F1, F2 formant trajectories
- **Goal**: arbitrary sentence's F1, F2 formant trajectories can be predicted with good accuracy speaker independently (currently in Hungarian)
- **Hypothesis:** A statistical parametric TTF model may produce similar accuracy as automatic formant trackers (eg. Snack, Praat)

2. Material and methods

The model is based on the multispeaker parallel formant database (FDB) with precise manual corrections and a HMM-based formant trajectory predictor:

- **Speakers of FDB:** 5 female, 5 male (Hungarian adults)
- **Parallel corpus**: same 1900 phonetically balanced sentences / speaker
- **Sound symbols**: SAMPA
- **Formant data of the FDB**: F1 (blue), F2 (pink) and F3 (green) measured by Praat at 5 locations (10%, 25%, 50%, 75%, 90%) of the sound and corrected manually in all vowels and m,n,J,j,l,v consonants (see example: figure top right)
- **Number of measured formants**: F1, F2, F3 altogether 7,125,000 values
- **Not measured, but linearly interpolated resonances of the oral vocal tract**: in voiceless sounds, and in voiced consonants, having weak formants: p,t,k,t',b,d,g,d',h,f,s,z,ts,dz,S,Z,tS,dZ,r; altogether 4,502,100 points (white)

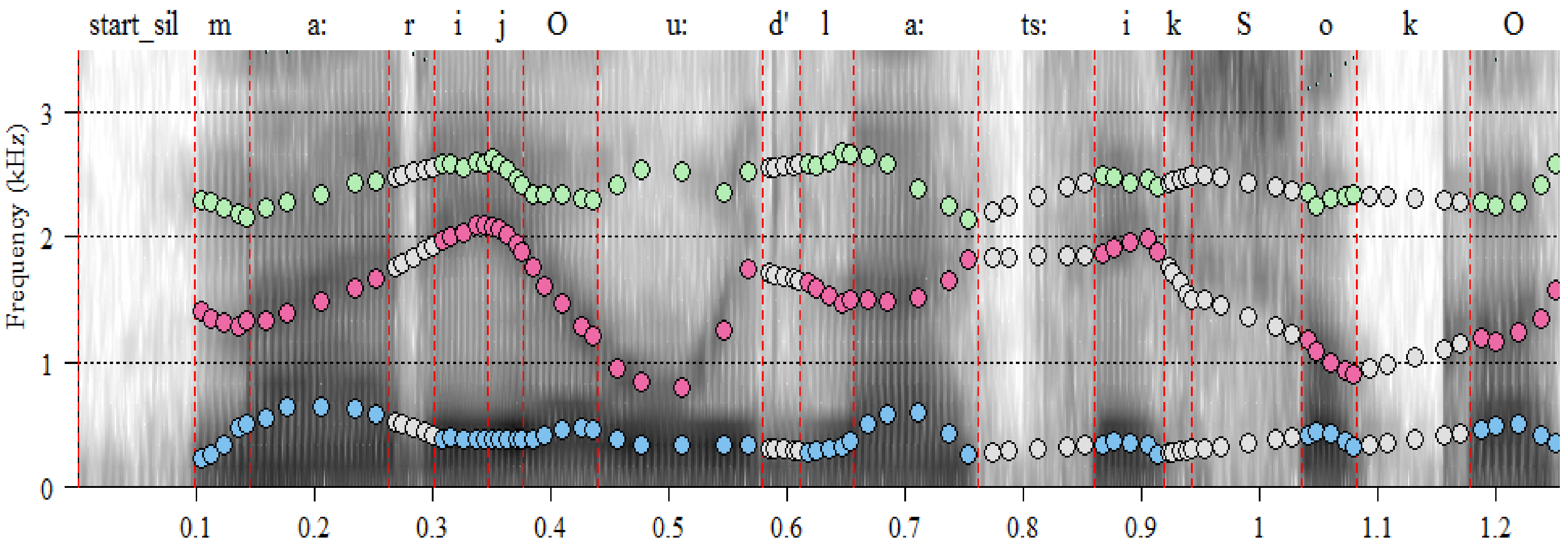

Figure 1. *Sample pattern from FDB for F1, F2 and F3*

The FDB was divided into 2 parts:

- Training database as corpus for training (90% of FDB) and
- Verification database (VDB) as corpus for verification (10% of FDB)

HMM-based formant trajectory predictor:

- The training of the HMMs was done with the HTS toolkit (version 2.2)
- Only F1, F2 from FDB-90% were trained as the goal was to build speaker independent (average) models and F3 is known to be speaker dependent

3. Generated formant trajectories

We trained the TTF models with various numbers of speakers as training data:

- **5sp.f, 5sp.m**, two models trained with 5 female/5 male speakers
- **1sp.f, 1sp.m** 10 models trained with 5 individual female/5 male speakers

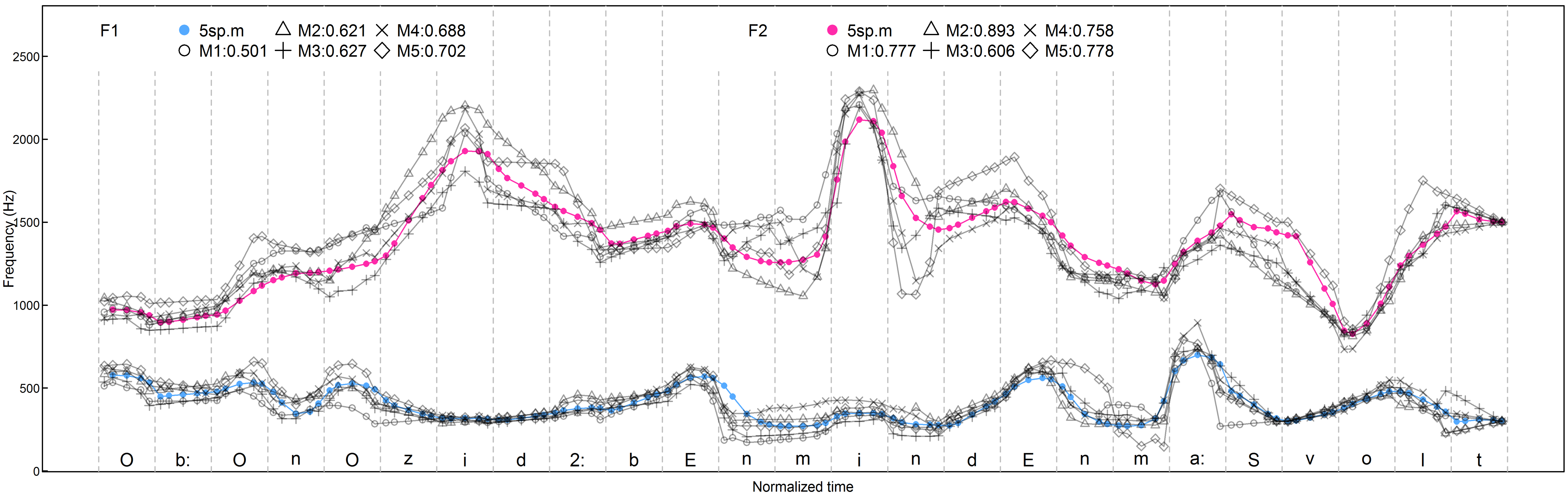

Figure 2. *Sample result of 5sp.m F1 (blue) and F2 (red) patterns. This graphical form is called „sentence formant pattern". The gray lines are the formant data of natural sentences. TMR values (the correlation between the 5sp.m TTF model and natural sentences) can be seen on the top.*

4. Evaluation

For the evaluation we introduce a Trajectory Matching Rate (TMR) which is based on the use of the correlation coefficient.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$TMR_j^s = r\left(\hat{F}_j^{N,s}, F_j^{N,s}\right),\ j = 1,2; s = Vowel, Cons$$

- $\hat{F}_j^{N,s}$ represents the normalized formant data produced by the TTF model
- $F_j^{N,s}$ denotes the normalized formant values of the same natural sentence of the VDB after determining *j* and *s. n* is the overall number of formant values

Properties of TMR:

- Its value is between –1 and +1
- The more similar the predicted sentence pattern (see upper) to that of the natural sentence of VDB, the closer its TMR value is to +1
- Scope of TMR is to compare sentences with same sound pattern

The validation was performed using the sentences in VDB:

- 10×190 sentences, not included in the training
- The predicted data produced by the model were compared with the original formant patterns of the 10 speakers of the VDB, sentence by sentence

Models	Praat	Snack	1sp	1sp	5sp
Compare with	VDB	VDB	same speakers from VDB	4 other speakers from VDB	VDB
Model/Formant tracker (as x-axis of Figure 3)	Praat	Snack	1sp	1sp*	5sp

5. Results

The main results for male and female data are shown on the figure below.

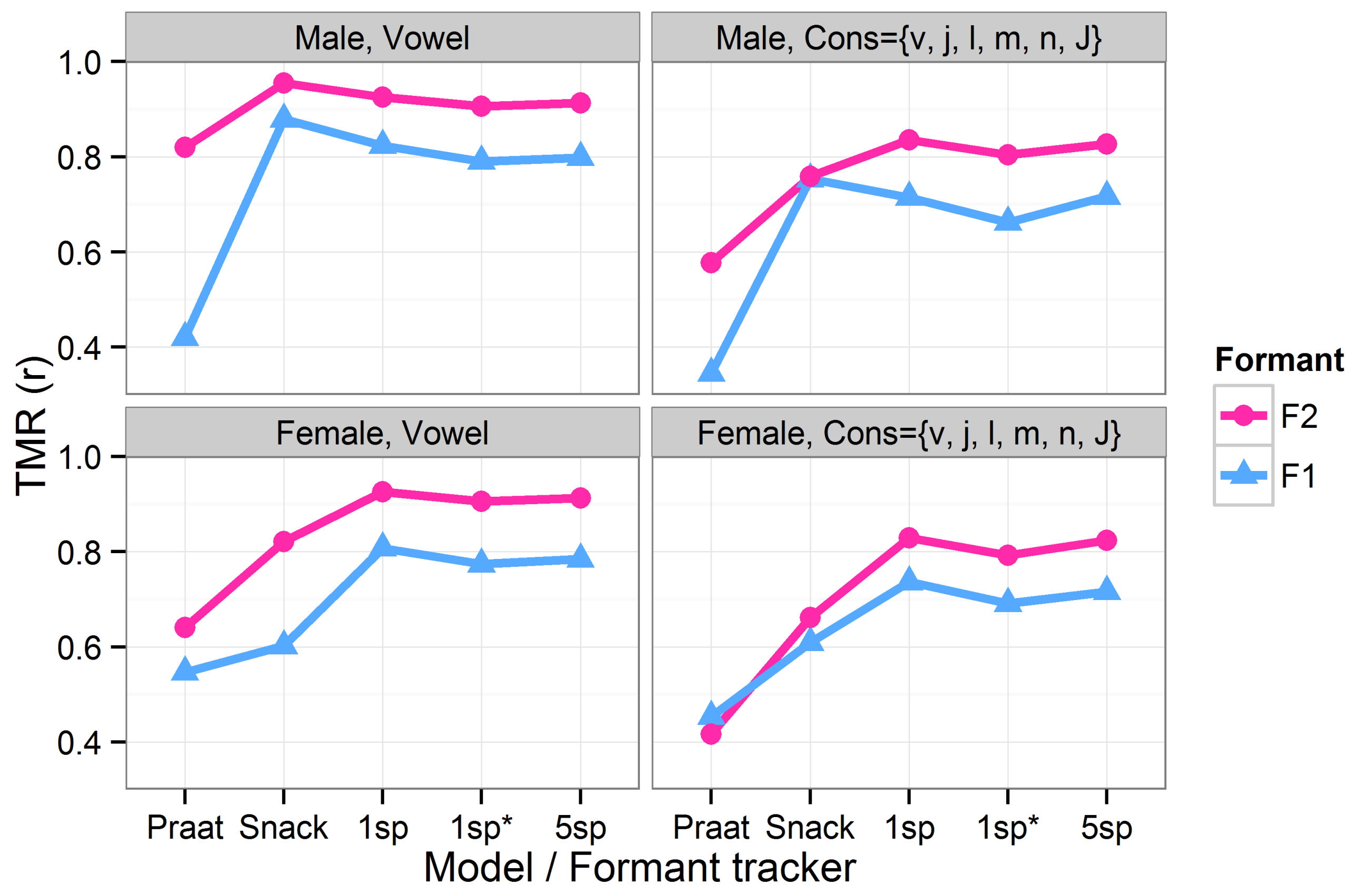

Figure 3. *Average TMR values for the different groups*

- Means of the Model / Formant tracker's TMR showed decreasing order as follows: **1sp** - 0.825, **5sp** - 0.812, **1sp*** - 0.791; **Snack** - 0.755, **Praat** - 0.527
- F2 can be predicted better (Mean: 0.867) than F1 (M: 0.751)
- The gender was not significantly related to the TMR values: means of male 0.810, female with the same conditions 0.808
- Vowels can be predicted better (M: 0.856) than v, j, l, m, n and J (M: 0.762)

6. Conclusions

- For 5sp models the hypothesis has been confirmed
- Mass formant prediction can be done directly from text using the TTF model
- Language specific calculations can be performed on formant trajectories
- Connecting with ASR, new ways of processing may be developed
- The method can be adapted to other languages as well

Live demo: http://hungarianspeech.tmit.bme.hu/ttf

The authors would like to thank the support of Swiss National Science Foundation via their joint research project (SCOPES Scheme) SP2: SCOPES project on speech prosody.