22nd Czech-German Workshop
on Speech Communication

Prague, Czech Republic
September 25-27, 2014

Quality Assessment in Pronunciation Trainer for Speech Disorder Therapy

I. Kraljevski[1], R. Jäckel[2], R. Kompe[1], G. Strecha[2],
F. Kurnot[1], M. Rudolph[1], D. Hirschfeld[1], R. Hoffmann[2]

[1]voice INTER connect GmbH
[2]Chair for System Theory and Speech Technology, TUD

TECHNISCHE UNIVERSITÄT DRESDEN

... smart signal processing from Germany

01 Introduction

- *Speech disorders - difficulties in articulation and intonation.*

- *Difficulties to control the larynx and the vocal cords (Hypokinetc Dysarthria), results in:*
  - *lower speech rate,*
  - *differently pronounced segments,*
  - *less consistent pronunciation for longer segments,*
  - *varying pronunciation due to fatigue.*

- *Speech therapy can improve production skills and the speech intelligibility.*

02 Speech Corpus

- *Acoustic analysis of oral production and pronunciation errors in speech impaired patients.*
- *Test and evaluation of automatic speech processing modules and the quality scoring concept.*
- *German native speakers: reference (1 M, 1F), 12 control (C) and 40 test (T) group.*
- *List of words and sentences provided by linguistics experts.*

03 Speech Characteristics of Hypokinetic Dysarthria

a) Respiration

-limited respiratory volume and low subglottal air pressure, reduced range of movement in respiratory muscles result in shallow breath support, poorly controlled exhalations and short breathing cycles; patients may have breathing rates faster than normal

b) Prosody

-monopitch, reduced stress and monoloudness

-inapproprite silences (due to akinesia)

-speech rate abnormalities (increased or decreased articulation rate)

c) Articulation

-imprecise consonants caused by reduced range of movements of articulators

-blurring of articulation;

-repeated phonemes usually at the beginning of an utterance or after a pause

d) Phonation

-harsh or breathy voice quality caused by incomplete vocal fold adduction (air leaking through a partly open glottis causes turbulent noise

e) Resonance

- patients may have increased nasal resonance caused by paresis of velum and/or reduced oral resonance

04 Acoustic data relating to articulation

(1) data relating to articulation of vowels:

Formant frequencies F1, F2 of 'corner vowels' ([i:], [u:], [a:]) allow for calculating VAI (vowel articulation index)

$$VAI = (F2/i/ + F1/a/)/(F1/i/ + F1/u/ + F2/u/ + F2/a/)$$

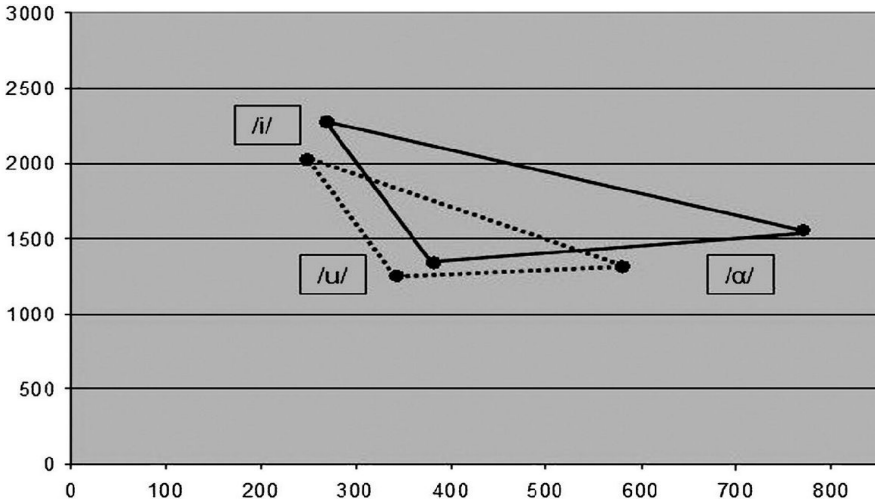


Fig. 01 Triangular vowel space of a healthy speaker (solid line) and a speaker with hypokinetic dasarthria (dotted line)

Source: Skodda et al.: Vowel articulation in Parkinson's disease, 2011

04 Acoustic data relating to articulation

- *As a consequence of reduced amplitude of articulatory movements, patients with Idiopathic Parkinson Syndrome tend to produce centalized vowels resulting in lower VAI.*

mean F1 (u)	mean F2 (u)	mean F1 (i)	mean F2 (i)	mean F1 (a)	mean F2 (a)	VAI	Population	m/f
312	1162	268	2187	592	1159	0,96188	IPS	m
323	1061	258	2251	628	1106	1,04933	Controls	m
360	1082	328	2523	828	1382	1,06818	IPS	f
351	933	299	2575	855	1319	1,18501	Controls	f

Table 1: Mean formant values F1, F2 for [a:], [i:], [u:] and VAI per population

mean F1 (u)	mean F2 (u)	mean F1 (i)	mean F2 (i)	mean F1 (a)	mean F2 (a)	mean VAI	Subset	Treshold
325	1329	266	2080	610	1193	0,86483	1 (N=9)	VAI < 0,9
304	1168	266	2164	572	1141	0,95113	2 (N=12)	VAI >= 0,9 < 1
311	1030	272	2291	597	1152	1,04542	3 (N=12)	VAI >= 1

Table 2: Distribution of formant values and VAI within population of speakers with IPS, m (N=33)

05 Acoustic data relating to articulation and phonation

Spectral centre of gravity (CoG) and Dispersion of spectral energy for fortis fricatives [s], [f] allow for calculating FSI („fricative sharpnes index")

FSI = CoG [s] / Disp. [s] + CoG [f] / Disp. [f]

Population	CoG_mean [s]	Disp_mean [s]	CoG_mean [f]	Disp_mean [f]	FSI (/f/;/s/)
Controls (m)	6259	2531	5598	3539	**4,15720**
IPS (m)	5242	2679	4572	3816	**3,19849**
Controls (w)	6936	2491	5919	4216	**4,29676**
IPS (w)	6242	2665	5078	3762	**3,77988**

Tab. 3 Mean GoG, Dispersion of spectral energy ([s], [f]), an d FSI per population

Decreased CoG, increased dispersion of spectral energy in fortis fricatives, and decreased FSI may indicate imprecise articulation and/or glottal palsy.

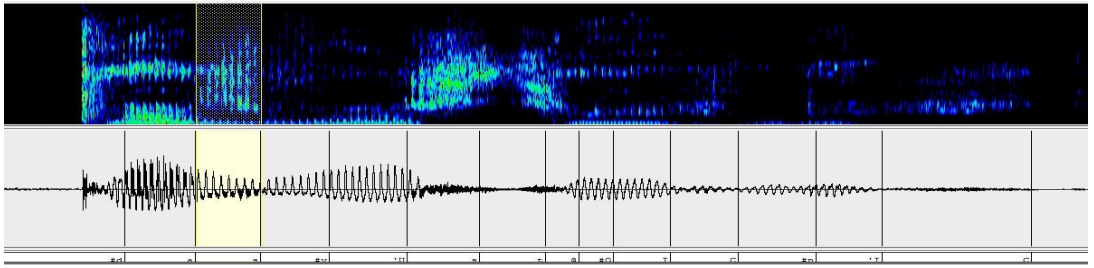


Fig. 2 Fully voiced segment /s/ in final position („das wusste"), speaker with IPS

Speakers with IPS frequently substituted fortis obstruents by their lenis opposites.

05 Acoustic data relating to articulation and phonation

Lenition of Fortis Obstruents in Utterances of Patients with IPS

Segment /Position	mean Duration (ms)		fraction_of_voiceless frames (%)		minimum Intensity (dB)	
	IPS	Controls	IPS	Controls	IPS	Controls
#f	99,16	112,96	74,89%	85,60%	54,66	51,60
.f	100,24	117,94	57,96%	77,51%	56,24	53,21
f	103,99	117,24	75,14%	84,74%	53,95	51,73
.s	91,65	113,05	52,79%	84,96%	59,64	58,09
s	100,28	106,03	77,65%	86,26%	58,75	57,35
#k	110,51	121,97	71,24%	78,14%	48,16	41,52
.k	87,46	108,22	45,44%	76,29%	54,87	44,76
k	95,34	101,42	50,44%	72,72%	50,21	42,32
#t	85,08	96,48	65,10%	79,92%	50,11	42,83
.t	72,16	84,72	43,82%	68,91%	57,11	48,27
t	76,92	86,65	62,02%	76,84%	52,58	45,81

Tab. 4 Duration, fraction of voiceless frames, and minimum intensity of fortis obstruents [f], [s], [k], [t] per position (#t=initial, .t=medial, t=final). Comparision of mean values for patients with IPS and controls.

06 Acoustic data relating to symptoms of hypokinetic dysarthria

Duration of phone segment (dur, ms), voicing (Praat: fraction of locally unvoiced frames in selection, %) and intensity (mean, min, max, dB) of fortis obstruents were measured for all consonant segments.

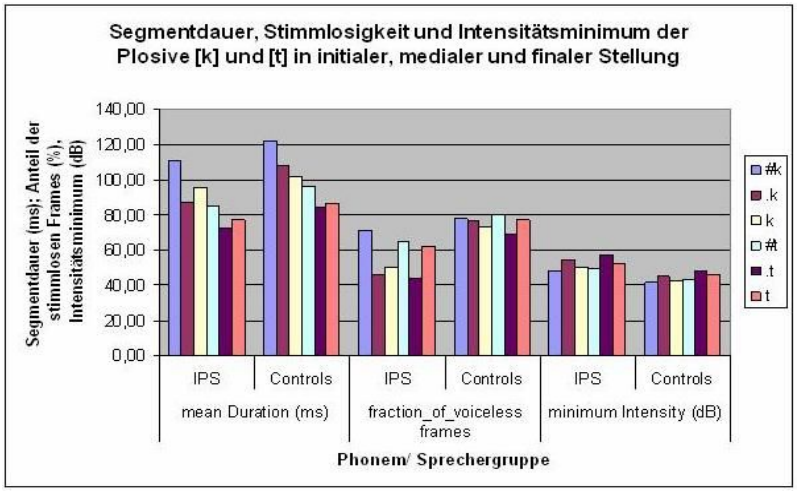


Fig. 04 Duration, fraction of voiceless frames and min. Int.ensity of [k], [t] in initial, medial and final position.

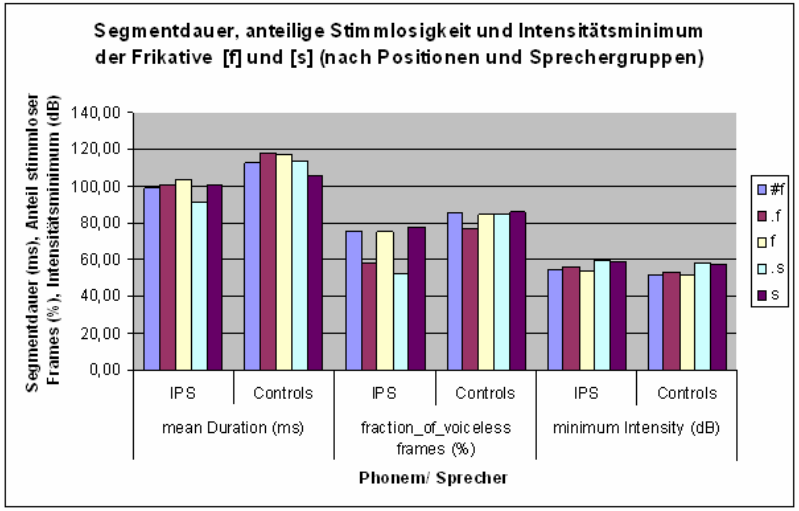


Fig. 05 Duration, fraction of voiceless frames and min. Int.ensity of [f], [s in initial, medial and final position.

Most evident differences have been observed in parameter „fraction of locally unvoiced frames" (patients lose ability to produce unvoiced obstruents). Pathophysiology of hypokinetic dysarthria includes dysfunction of vocal fold kinematics (i.e., slow opening and inadequate closing of the vocal folds), vocal fold asymmetry and bowing,and vocal fold paresis).

07 Phonation, voice quality

Harsh or breathy voice quality is caused by incomplete vocal fold adduction (air leaking through a partly open glottis causes turbulent noise)

***Roughness** (R), **breathiness** (B) and **hoarsness** (H) were evaluated basing on speech samples according to 3 speech modalities: (1) repeating, (2) reading, (3) picture description*

Population	m/w	mean B	mean R	mean H
IPS	m	0,9848	0,7879	1,1667
IPS	f	0,7857	0,3929	0,8929

Tab. 5 Results of RBH rating (expert rating, carried out by B.J. Kroeger, RWTH Aachen) RBH rating uses discrete values: 0 = non; 1 = mild; 2 = moderate; 3 = sever.

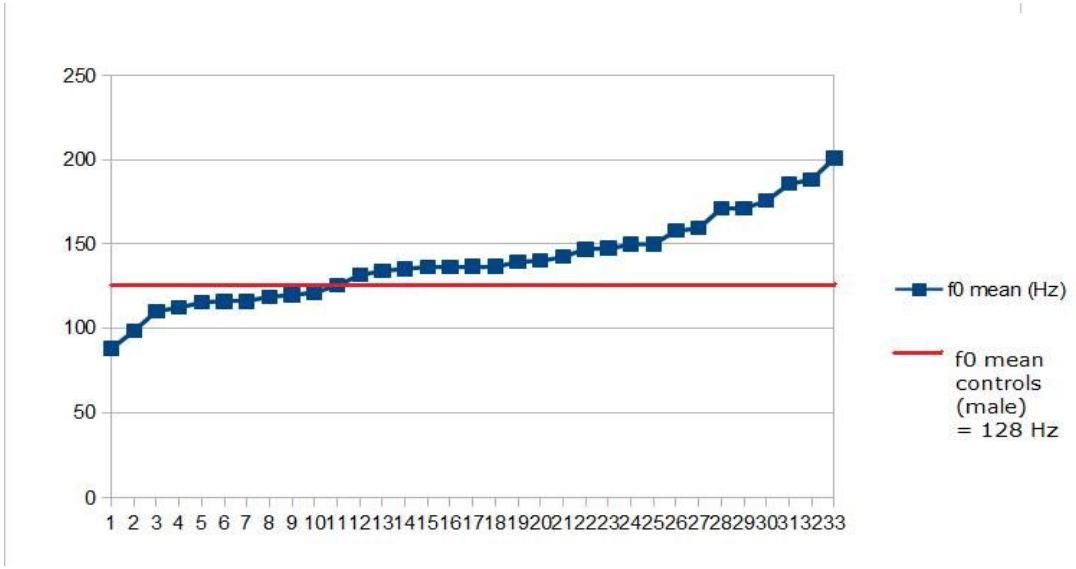
Jitter, Shimmer and HNR were measured on prolonged vowels (Praat voice analysis)

Population	Jitter local	Shimmer loc.	HNR
IPS	0,918	5,975	18,875
Controls	0,744	4,200	20,552

Tab. 6 Mean values of jitter (local), shimmer (local) and Harmonics to Noise Ratio for Patients with IPS and controls. Tresholds of pathology: Jitter: > 0,820 Hz; > Shimmer: 4,85 dB; HNR: < 20 dB

08 Mean f0 and f0 SD for patients with IPS and controls

Population	f/m	F0 min (Hz)	F0 max (Hz)	F0 range (Hz)	F0 mean (Hz)	F0_SD (% F0)
Controls	f	70	336	265	179	19
IPS	f	69	333	263	179	20
Controls	m	71	249	178	128	22
IPS	m	73	291	218	**141**	**16**

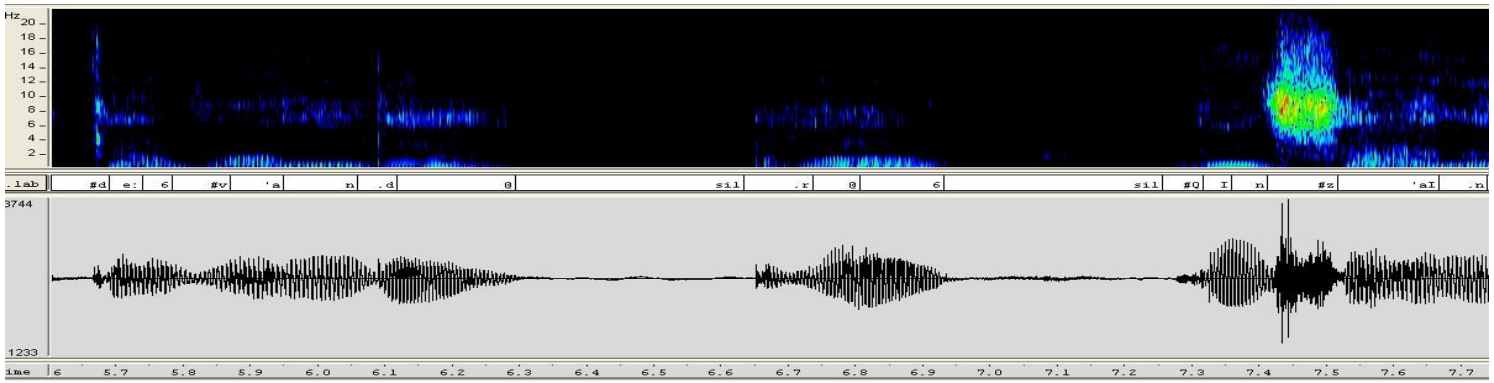


Male speakers with IPS had significantly increased f0 mean and decreased f0 SD values. Two third of male speakers featured mean f0 exceeding the average f0 of healthy controls (128 Hz).

09 Temporal characteristics: LAR and Pauses to Phonation Ratio

Patients with IPS may exhibit both increased and decreased articulation rate. The average LAR of all patients exceeds average LAR of controls. 20 out of 40 patients with IPS had mean LAR values above 13 phones/sec., 10 of them above 14 phones/sec.

Population	mean_dur_SG (ms)	mean_LAR (phones/sec.)	mean Ratio Pauses/Speech
Controls	78,90	12,86	8/89
	10,50	1,74	0,04
IPS	74,71	13,73	7/52
	12,77	2,14	0,07



The main reason for higher pauses to phonation ratio is short breathing cycles.

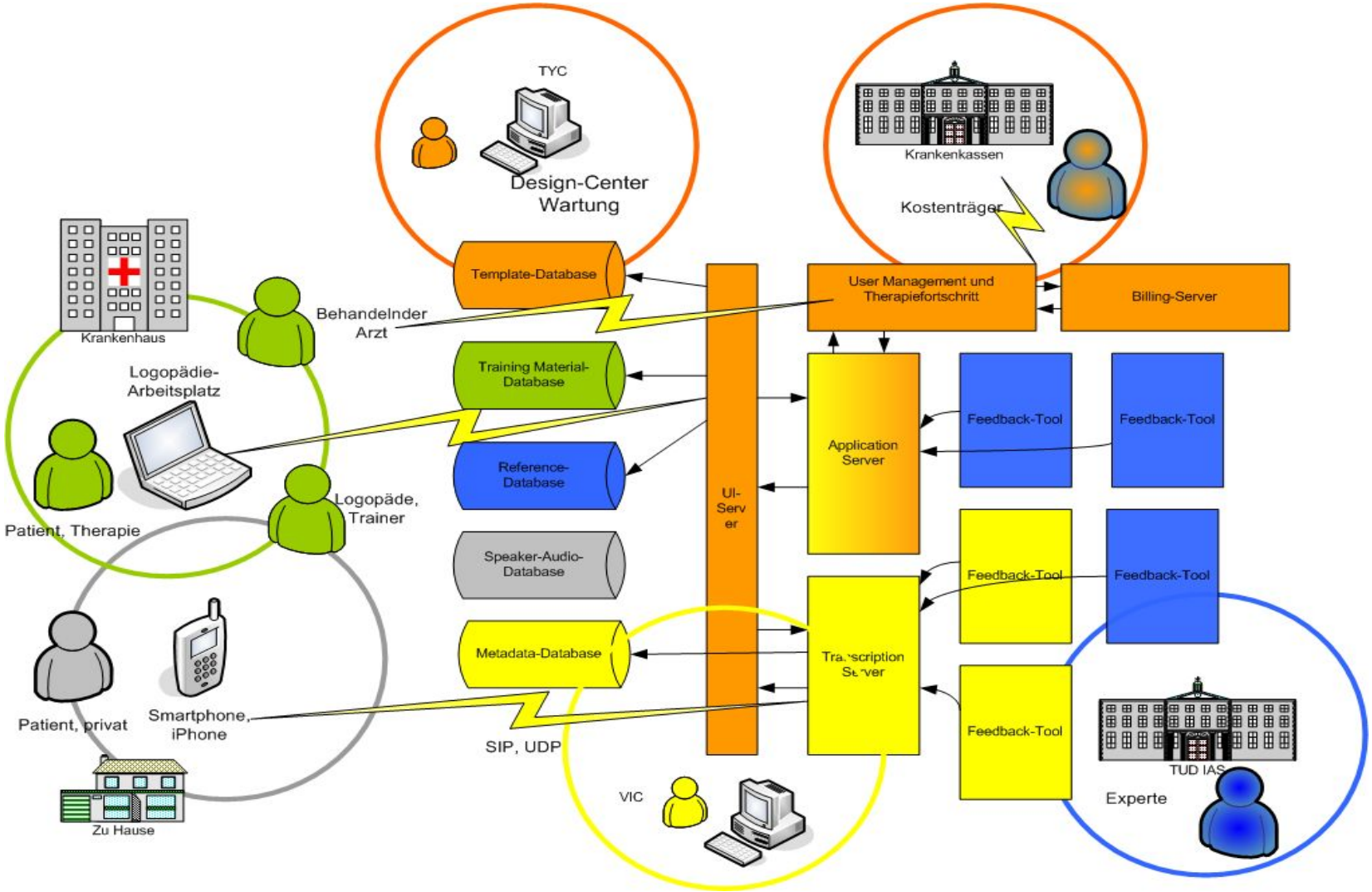
10 Computer Assisted Pronunciation Training

Computer Assisted Pronunciation Training (CAPT):

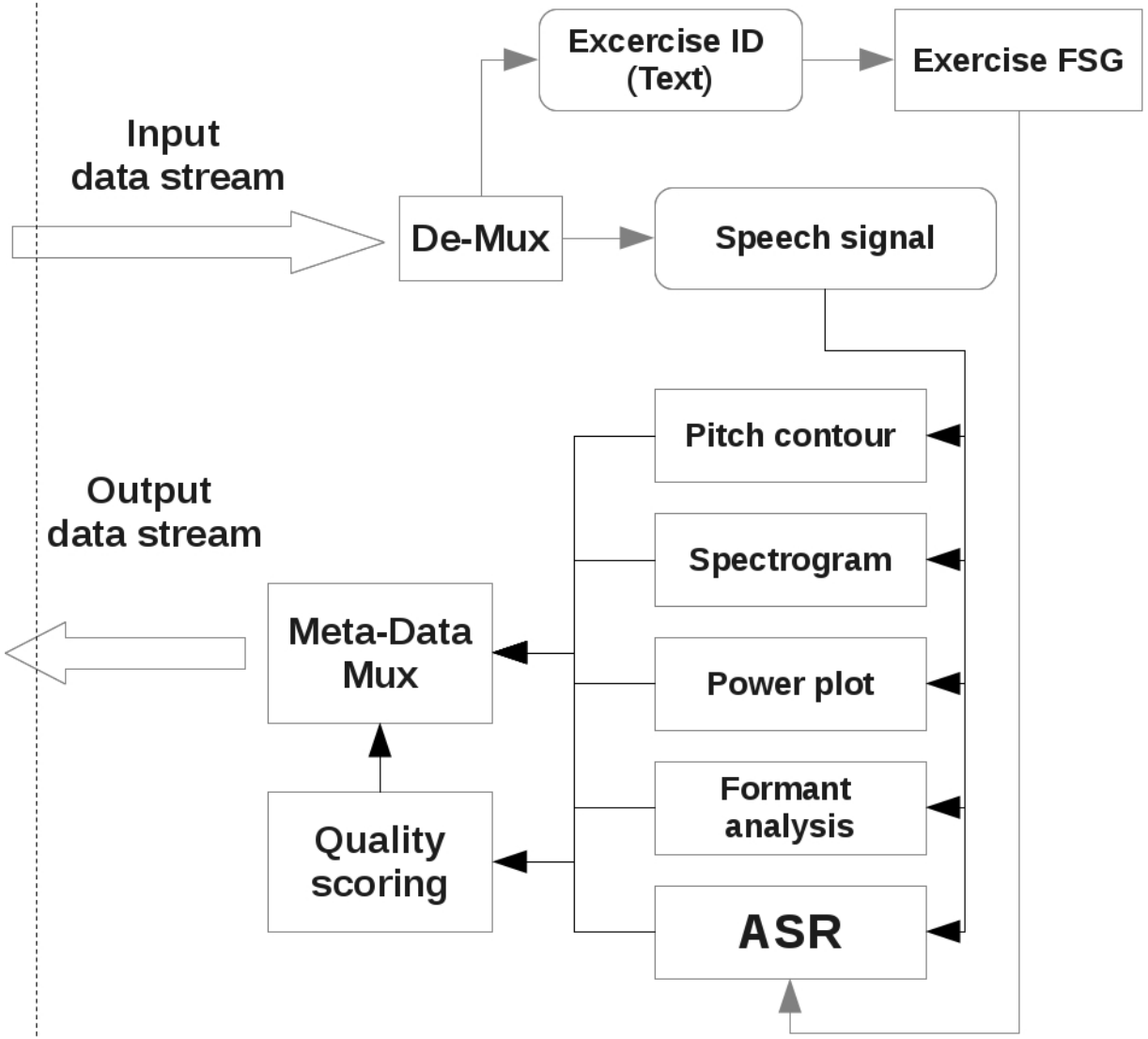
- *Client-server architecture*
- *Real-time processing and results analysis.*
- *Content sharing by community based multimedia database.*

- ***Exercises: words, word-pairs, phrases, longer texts***
  - *Breathing*
  - *Loudness level*
  - *Pitch range*
  - *Prosody / intonation / melody*
  - *Articulation of phonemes*

10 Computer Assisted Pronunciation Training



11 Pronunciation Trainer



11 Pronunciation Trainer

- *Pitch and formant contours:*
  - *LPC cepstrum (f0) and LPC spectra peak picking*
  - *Median filtering*

- *Phoneme segmentation:*
  - *Pocketsphinx ASR engine (Gstreamer plug-in)*
  - *AM trained with Phondat I and Verbmobil I.*

- *Phoneme recognition - lexicon consist only of phonemes.*

- *Separate FSG for each exercise.*

12 Pronunciation Scoring

- *Forced-alignment – user's compared with reference utterance*

- *Acoustic (ACS_n) and duration (Dp_n) scores were estimated for each pair of user and reference phoneme:*

$$ACS_n = 1 - \left| \frac{ACS_{ref} - ACS}{ACS_{ref}} \right| \qquad Dp_n = 1 - \left| \frac{Dp_{ref} - Dp}{Dp_{ref}} \right|$$

$$GOP = \alpha \cdot ACS_n + (1 - \alpha) \cdot Dp_n$$

12 Pronunciation Scoring

- *Contour matching against the reference utterance*
  - *Dynamic Time Warping (DTW)*
  - *Root Mean Square Error (RMSE)*

Weighted linear combination of the RMSE, deletions and insertions:

$$\text{Quality Score} = (11 - \alpha * \text{RMSE} + \beta * D + \gamma * I) / 5$$

Where α, β, γ are the weighting coefficients with their sum equal to 1.

13 Corpus Analysis

- *Automatic corpus analysis with the CAPT system.*
- *Quality scores were estimated for 40 T and 7 C subjects on exercise set of 34 sentences.*
- *Statistical analysis - dependencies and the contribution to the final score.*
- *Description of the overall quality including the articulation and the intonation.*

13 Corpus Analysis

Table 1 - Wilcoxon Mann-Whitney Test between the control (N=238), and test group (N=1426),

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Score	P-value (95%) C-T <> 0	P-value (95%) C-T > 0	Significance
ST Energy	0,0001	0,0001	***
Spectrogram	0,0000	0,0000	***
F0	0,0323	0,0161	*
Formants	0,0002	0,0001	***

Table 2 – Correlation matrix

	ST Energy	Spectrogram	F0	Formants
ST Energy	1.0000	0.7015	0.5460	0.6465
Spectrogram	0.7015	1.0000	0.6469	0.7444
F0	0.5460	0.6469	1.000	0.6305
Formants	0.6465	0.7444	0.6305	1.0000

13 Corpus Analysis

Phoneme articulation

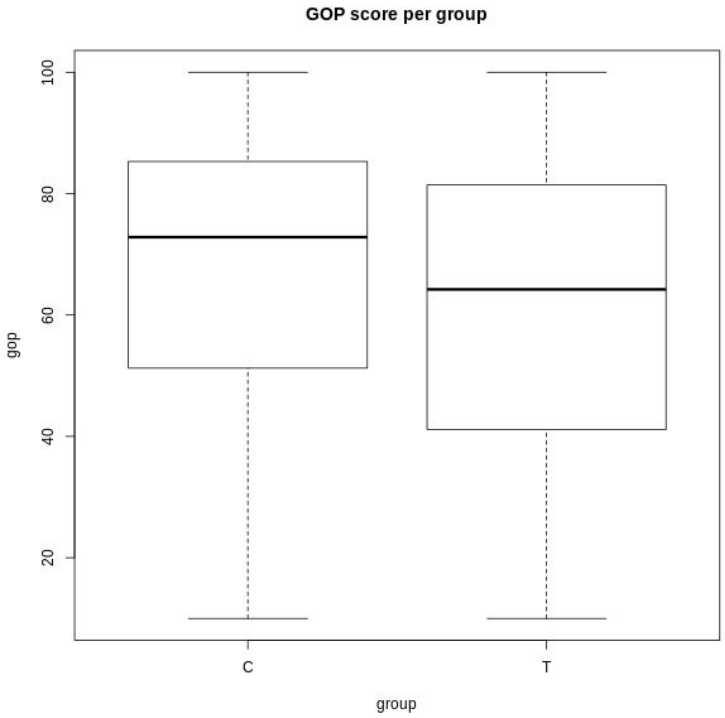
- *23318 occurrences (3363 control and 19955 test).*

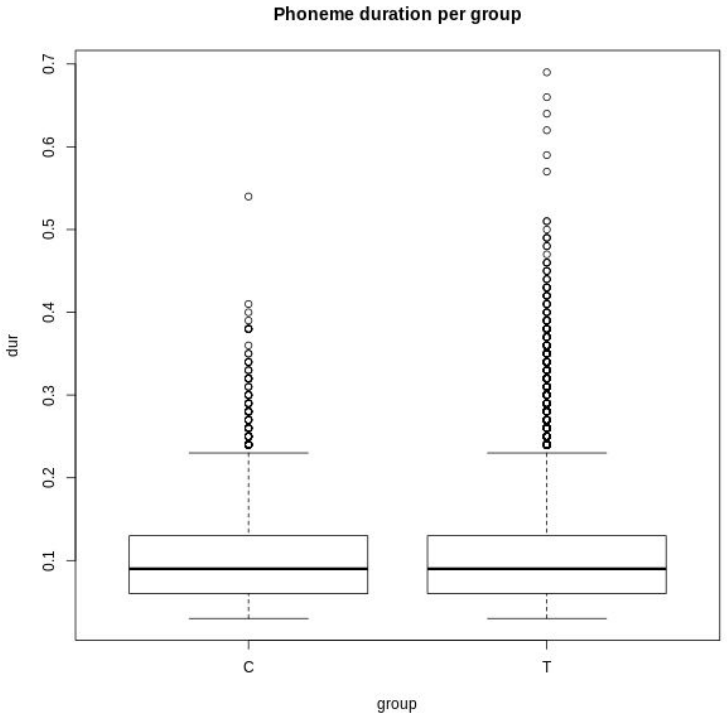
Table 3 - Wilcoxon Mann-Whitney Test for GOP scores and phoneme duration per group

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Score	P-value (95%) C <> T	P-value (95%) C < T	P-value (95%) C > T	Significance
GOP	0,0000	1,0000	0,0000	***
Duration	0,1296	0,9352	0,0648	

13 Corpus Analysis





- *Cross-group comparison - GOP scores and phoneme durations.*
- *The control group achieved better results (higher mean, low deviation, less outliers) than the test group.*

14 Acoustic characteristics and quality scoring

- *The following acoustic parameters were compared with the averaged GOP score per speaker:*
  - *VAI – vocal articulation index ([a:], [i:], [u:])*
  - *FrAI – fricative articulation index ([f], [s], [C])*
  - *FSI – Fricative sharpness index ([f], [s])*

- *Pearson correlation (95% confidence):*
  - *corr(VAI, gop([i:]))= 0.6294 moderate*
  - *corr(VAI, gop([u:]))= 0.4672 weak*
  - *corr(VAI, gop([a:]))= 0.3193 weak*
  - *corr(FrAI, gop([s]))= 0.6928 moderate*
  - *corr(FrAI, gop([f]))= 0.4913 weak*
  - *corr(FrAI, gop([C]))= 0.2359 no linear relationship*
  - *corr(FSI, gop([f]))= 0.5330 moderate*
  - *corr(FSI, gop([s]))= 0.7306 strong*

14 Acoustic characteristics and quality scoring

- *The regression analysis just confirmed that the parameters with high correlation can be used to model and predict the acoustic measurements from the automatically obtained quality scores:*
  - *VAI = f(gop([i:]), gop([u:])):* R^2_{adj}=0.4571,
  - *FrAI = f(gop([s])):* R^2_{adj}=0.4766
  - *ASIfs = f(gop([f]), gop([s])):* R^2_{adj}=0.5353

- *However the dependency is not so strong since the acoustic characteristic measures are absolute, and the automatically derived scores by the CAPT system are relative to a reference speaker.*

15 Conclusions

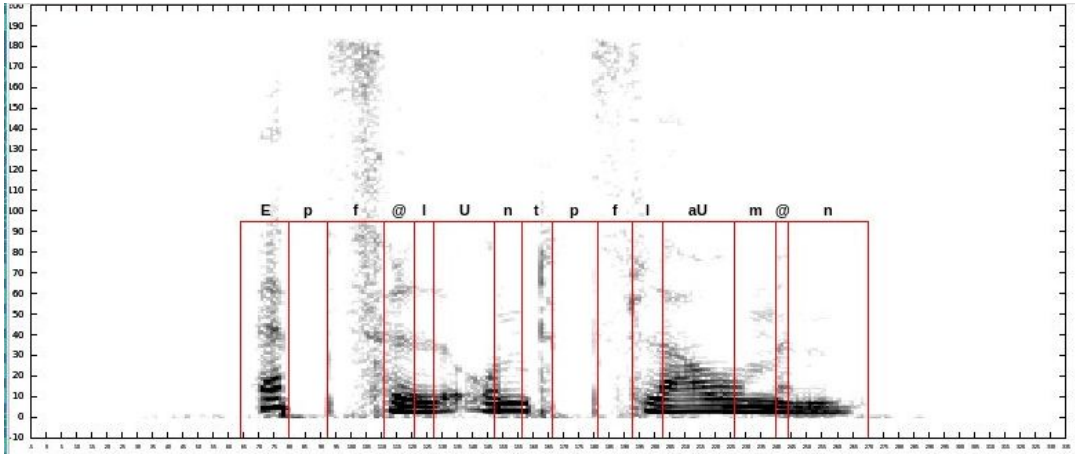
- *Computer-Aided Speech Therapy system*
- *Pronunciation and prosody are compared against reference speech.*
- *Speech processing in real-time.*
- *Statistical confirmation - scores significantly differ across speakers groups (control and test).*
- *Moderate correlation of: VAI with GOP(/i:/)*
- *Moderate to strong correlation of: FrAI and FSI with GOP(/s/).*

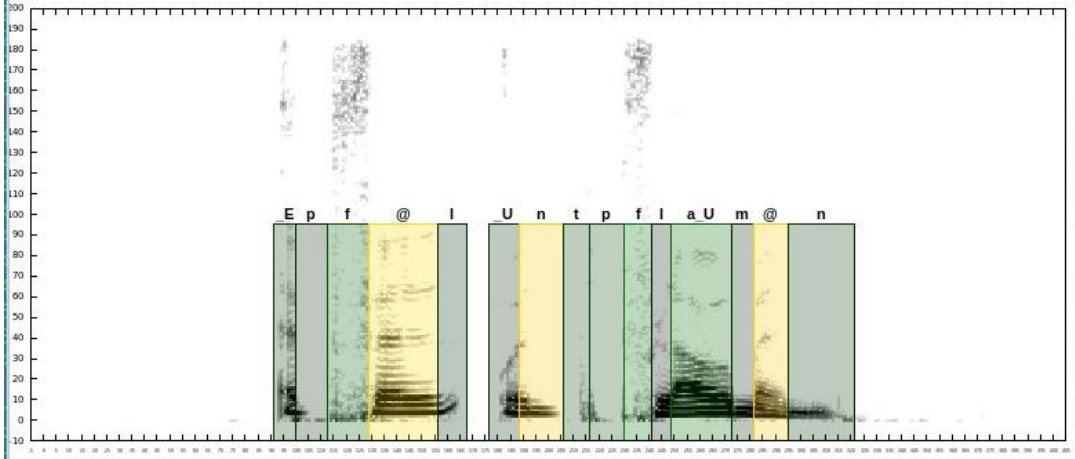
THANK YOU FOR YOUR ATTENTION

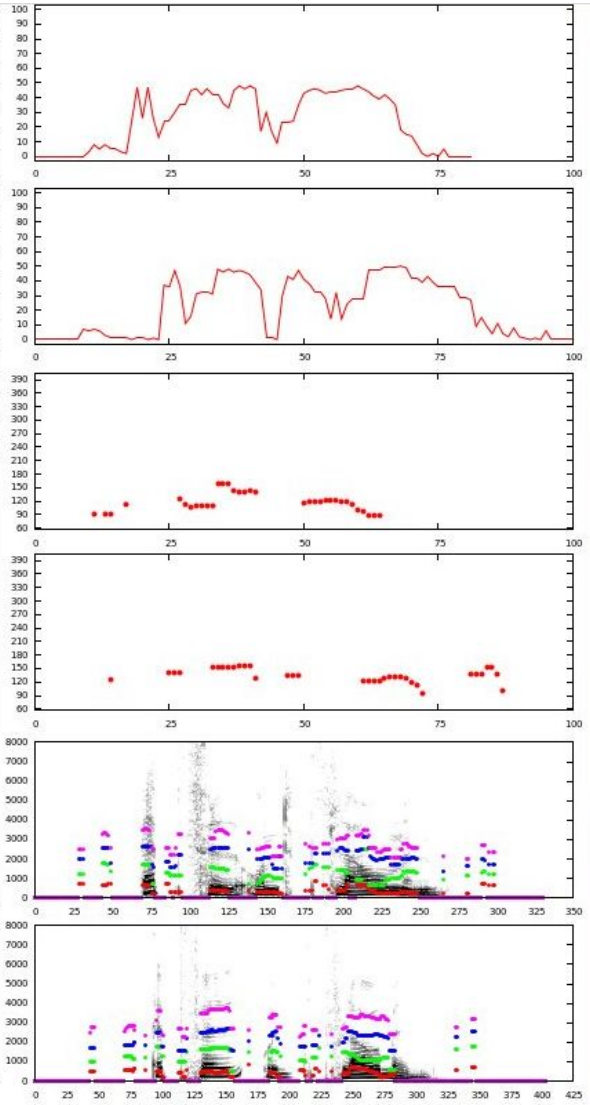
voice INTER connect

Voice INTER connect GmbH
Ammonstraße 35
01067 Dresden, Germany
Tel.: +49 (351) 407526 50
Fax.: +49 (351) 407526 55
www.voiceinterconnect.de

EXAMPLE







PlgID=1: 5.18
PlgID=2: 6.17
PlgID=3: 6.73
PlgID=4: 7.56

PlgID 1-Short Time Energy
PlgID 2-Spectrogram
PlgID 3-Pitch
PlgID 4-Formants
D-Deletions
I-Insertions
(Duration normalized values)