SpeD 2009

Constanța, Romania
June, 18-21, 2009

ASSESING THE QUALITY OF VOICE SYNTHESIZERS

Horia-Nicolai Teodorescu[*],[**]
Monica Feraru[*]
Marius Zbancioc[*],[**]

Institute of Computer Science
ICS
Iasi-Romania

[*] Institute for Computer Science, Romanian Academy
[**] Technical University of Iasi, Romania

UNIVERSITATEA TEHNICA "GHEORGHE ASACHI"
1813
IASI

Overview

- Introduction
- The methodology
- The analysis
- Results
  - Results of the assessment by listeners
  - Visual inspection of the spectrograms and pitch graphs
  - Analysis of the durations of the sounds
  - Results of the numerical quality assessment based on formantic features
- Discussion and Conclusions

AIM

1. Determine why some synthesizers sound so unnatural

2. Establish a method for quantitative assessment of voice quality

Introduction (I)

- Many research groups are trying to improve the quality of the concatenative synthesis voice.
- The popular method for assessing the quality of the synthetic voices is subjective and determined by a statistic score obtained on many listeners.
- Unnatural voice signals distort and hamper the understanding and the emotional response to voice communication.
- We propose a methodology based on comparison of human with synthetic voice.

The methodology (I)

- We used the (feminine) human voice from the SRoL corpus and the synthetic voice obtained with the BAUM™ and Ivona™ synthesizer.
- Investigated: 5 feminine voices which has fundamental frequency near to the fundamental frequency of the synthetic voice.
- The sentences were annotated, determined: the values of the formants (F1, F2, and F3);
- Computed: the ratios F1/F0, F2/F0, F3/F0, the average values, and the standard deviation for these ratios, the difference between the average of the ratios for human and for synthetic voice.

The methodology (II)

- Estimated:
  - the average values of the durations of all vowels and consonants for each person and for all speakers.
  - the standard deviation of the duration for all vowels and consonants for all speakers, on all analyzed sentences.
- Method based on ratios of values (formants vs. pitch); it is highly sensitive to erroneous measurements in any of these parameters.

The analysis (I)

- Causes for unnaturalness are:
  - poor concatenation from speech contexts that are inappropriate;
  - poor prosodic dynamics;
  - the unnatural (erroneous) relation between the formants and the pitch;
  - To fast transitions at the border of the concatenated segments.

The analysis (II)

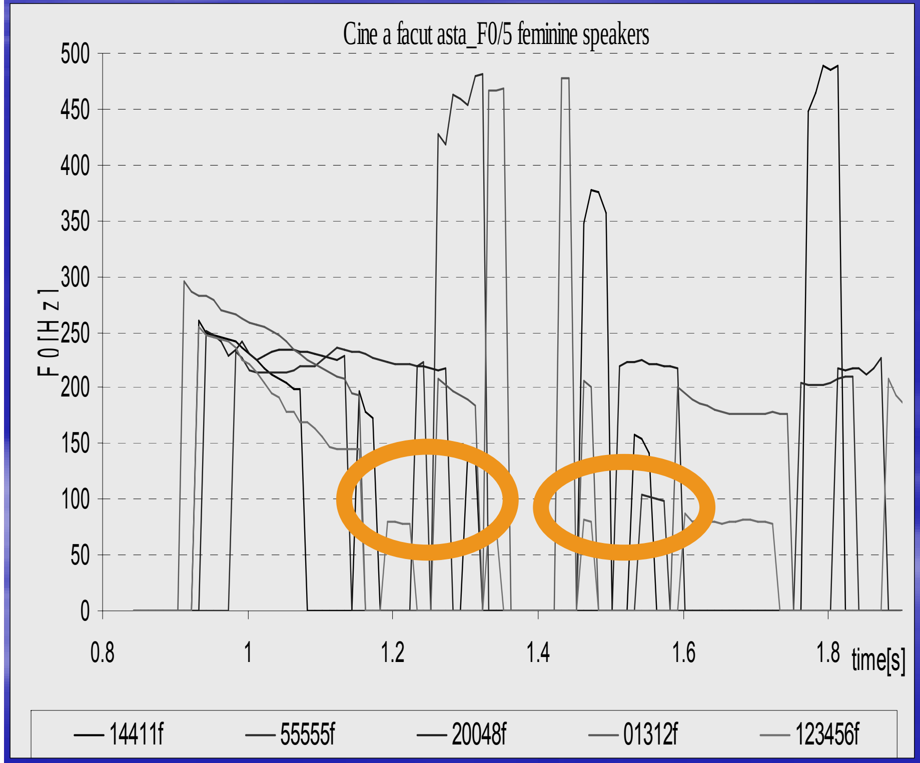
- The correction was made automatically, according to the algorithm:

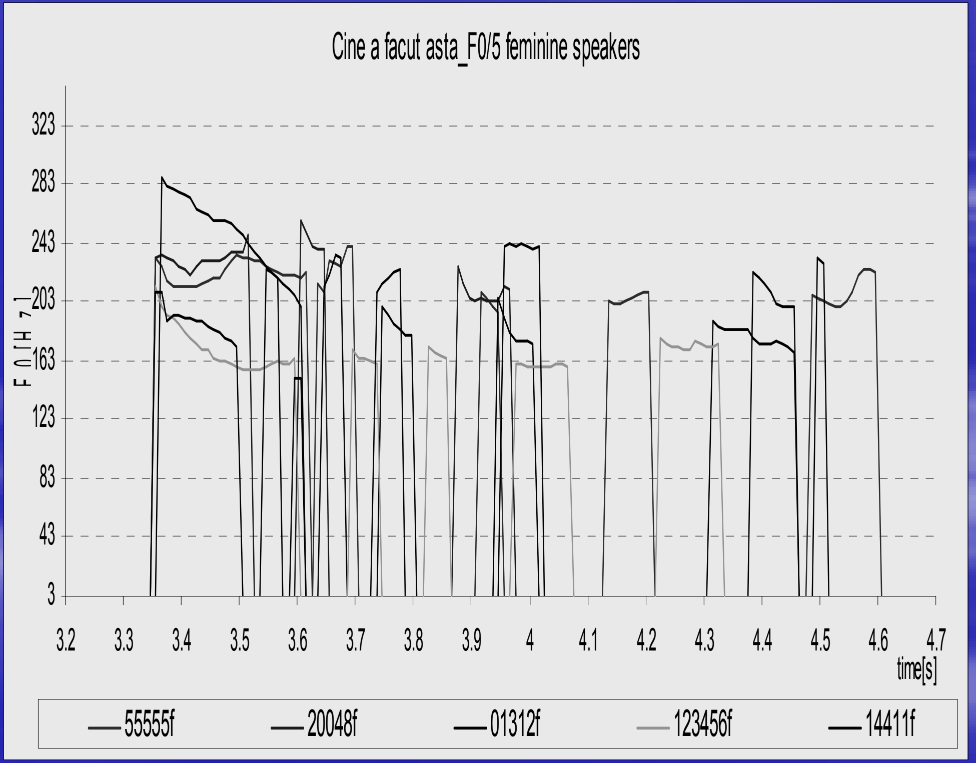
If $F_0(t) < 60\% \cdot average(F_0)$ but $F_0(t) > 35\% \cdot average(F_0)$,then $F_0(t) = 2F_0(t)$

If $F_0(t) > 150\% \cdot average(F_0)$, then $F_0(t) = F_0(t)/2$

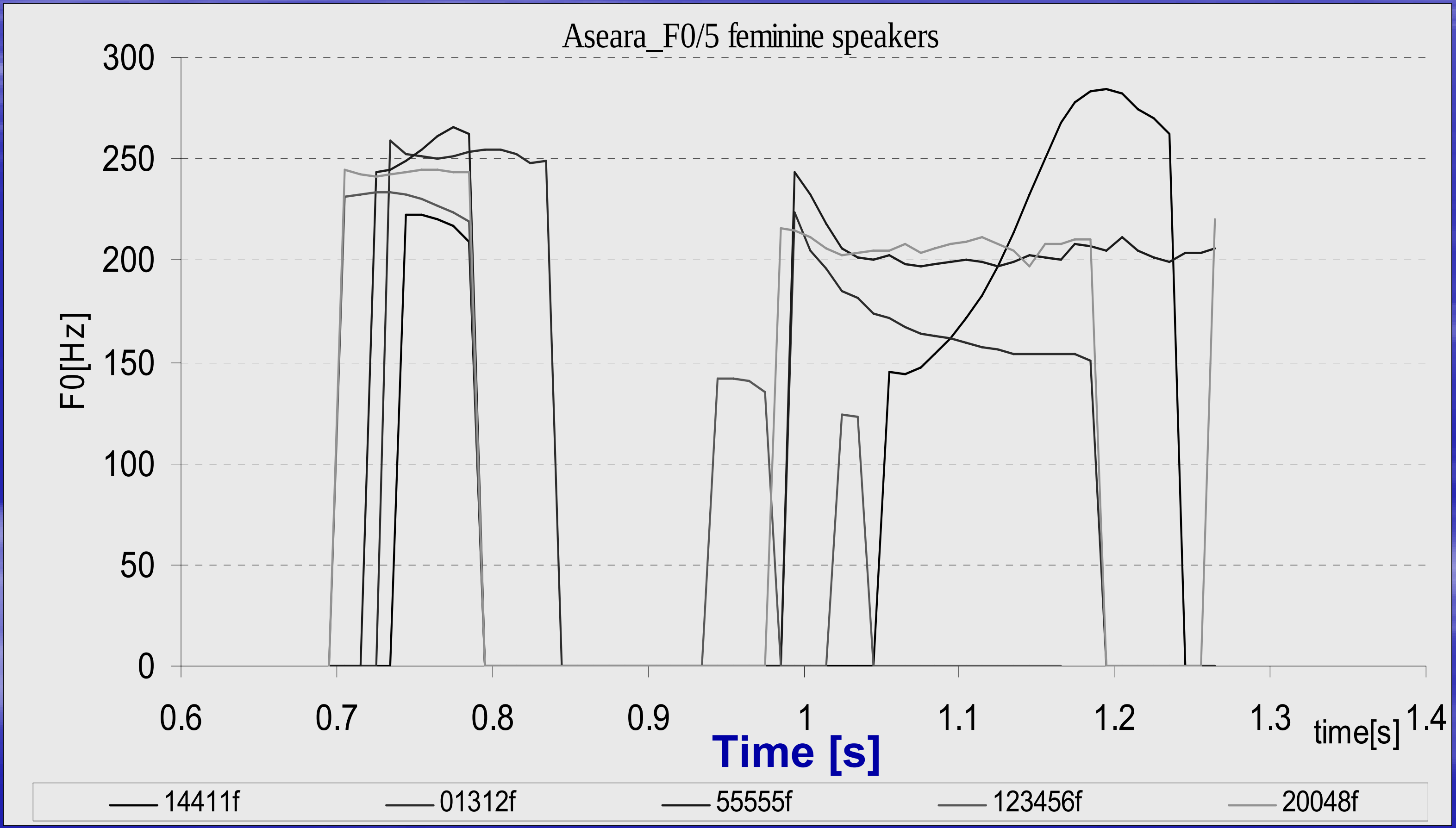
If $F_0(t) < 35\% \cdot average(F_0)$, then $F_0(t) = 3F_0(t)$

The values of F0, for 5 feminine speakers spelling with neutral tone the sentence “Cine a făcut asta?” (first pronunciation, left panel), and corrected values (right panel)





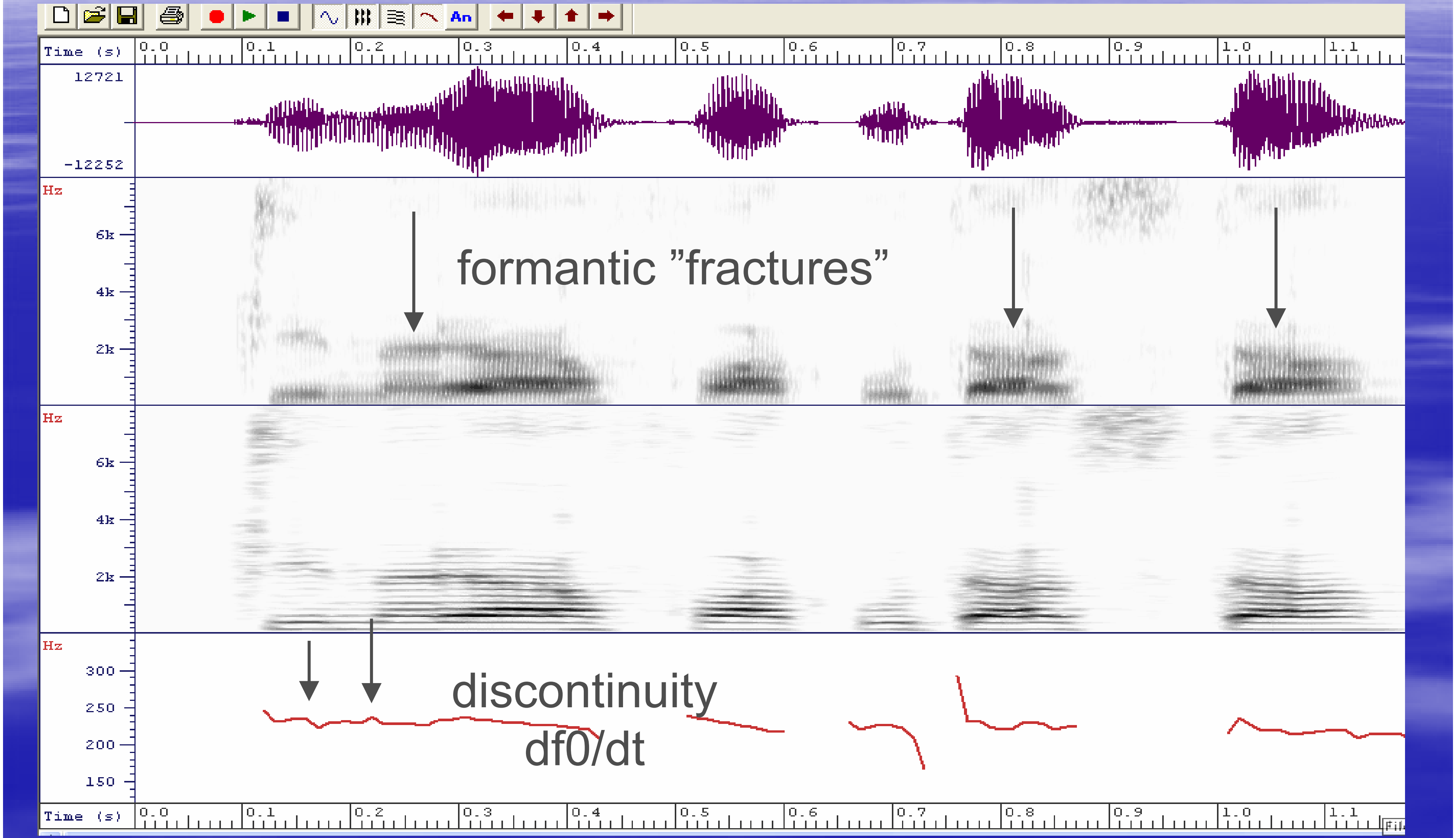
The values of F0, for 5 feminine speakers spelling with neutral tone the sentence Vine mama (first pronunciation), values given by Praat



Results of the assessment by listeners

- **Person 1**: the prosody is very poor; from all 4 sentences, “Cine a făcut asta?” is the worst pronounced; the word *asta* is the only which is well pronounced; the pauses between the words are too short.
- **Person 2**: the prosody is bad; *c* and *t* from the word *făcut* have a poor connection with the neighboring vowels; the absence of pauses gives the impression of a single word instead of a sentences; *i* is pronounced too short.
- **Person 3**: the beginning of the sentence is uttered in an unusual way; the prosody is missing; the recording is unpleasant for the hearing.

The evolution of pitch, using Wasp software; synthetic voice - sentence “Cine a făcut asta?”



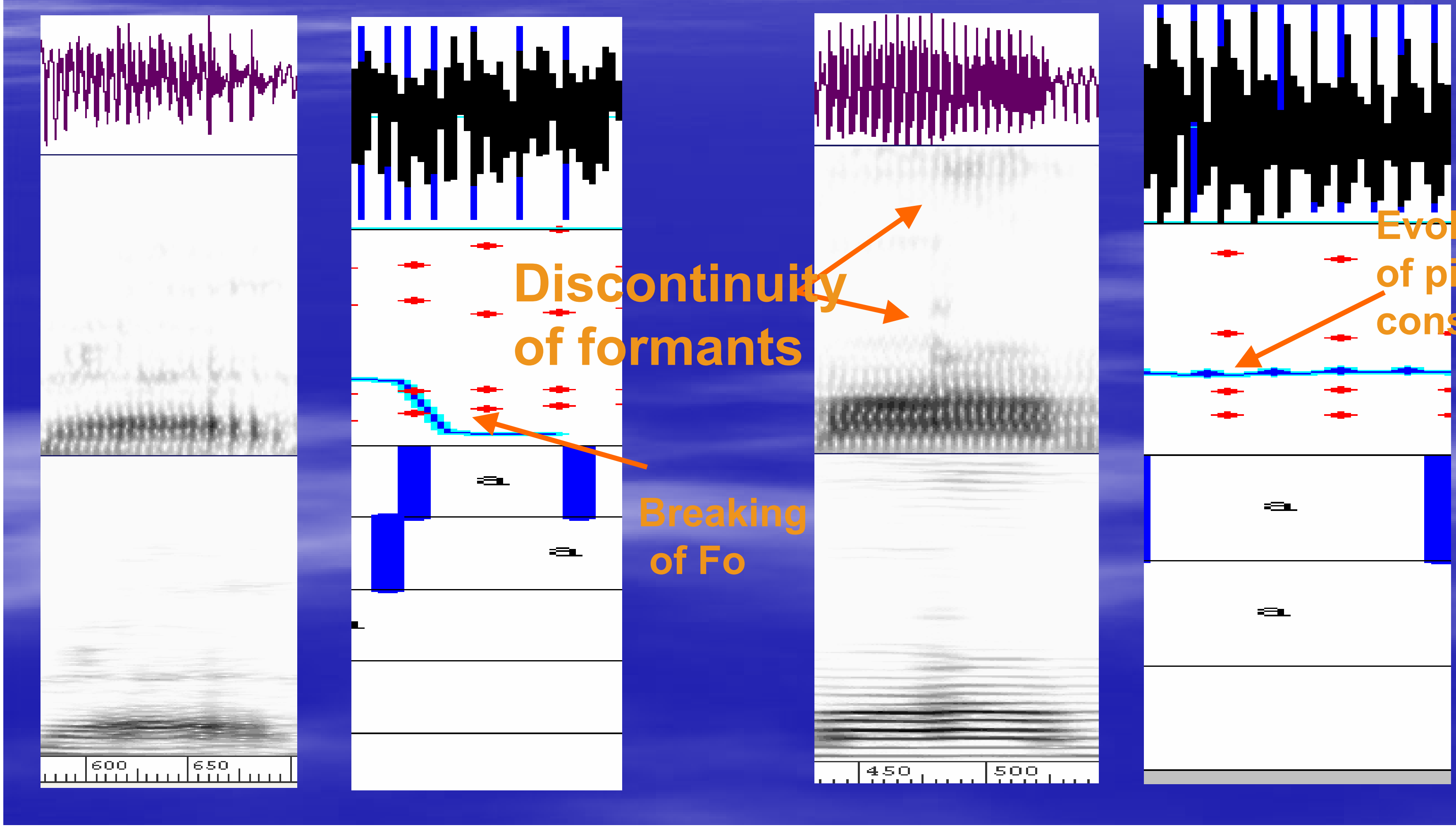
formantic ”fractures”

discontinuity df0/dt

Comparation between human and synthetic voice –selectated segment “a1” from “mama”, sentence “Vine mama”

human voice

synthetic voice

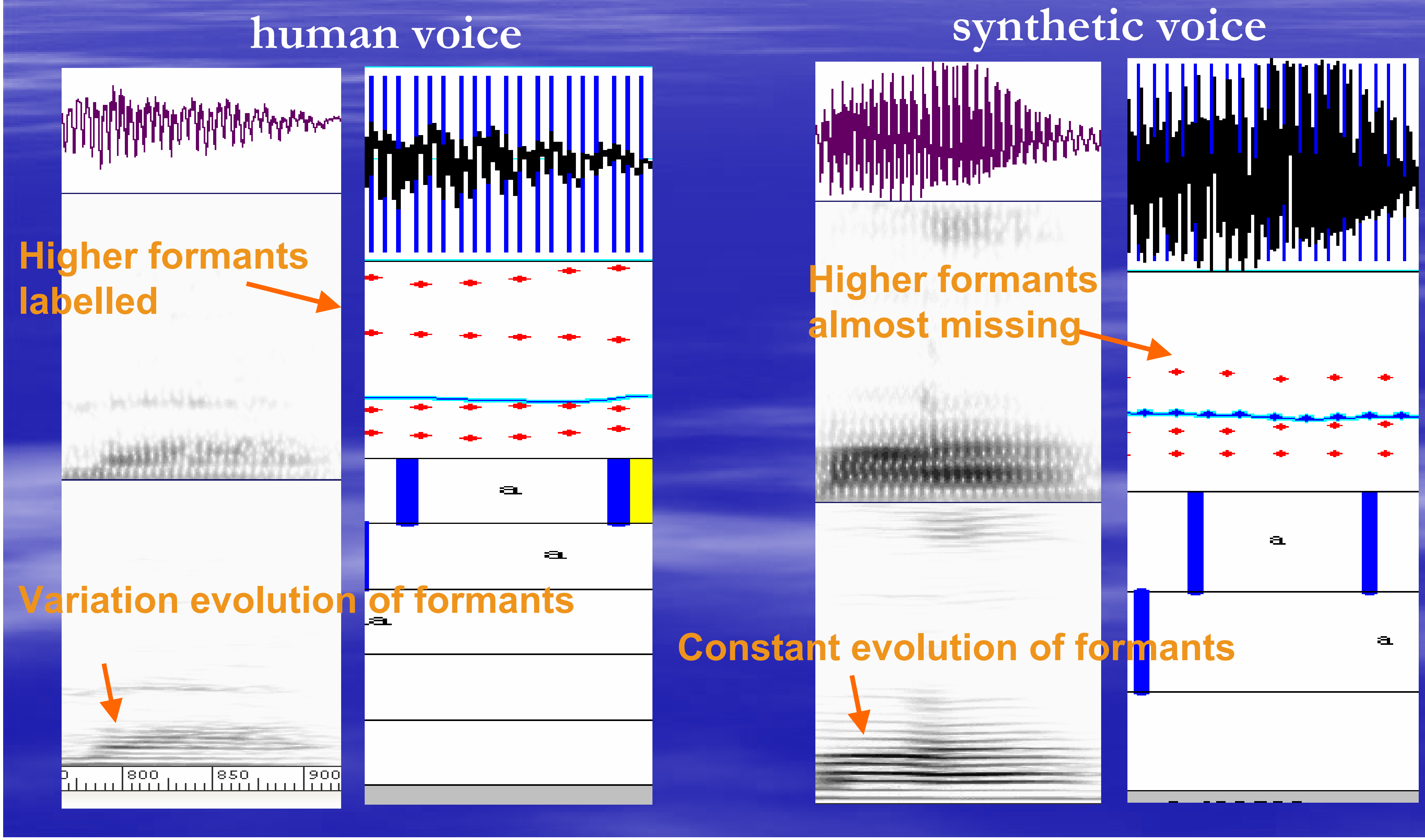


Discontinuity of formants

Breaking of Fo

Evolution of pitch is constant

Comparation between human and synthetic voice –selectated segment “a2” from “mam**a**”, sentence “Vine mama”



human voice

Higher formants labelled

Variation evolution of formants

synthetic voice

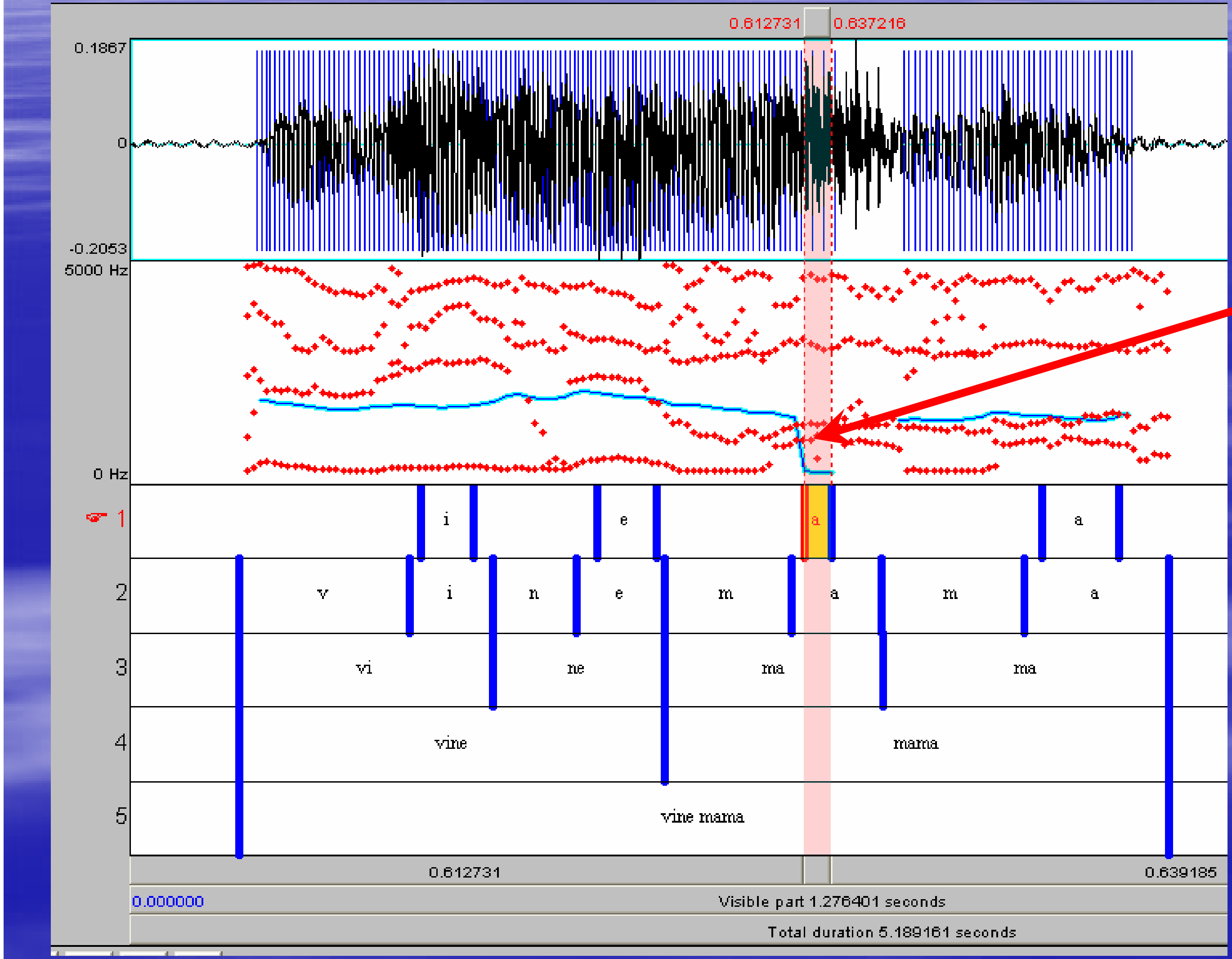
Higher formants almost missing

Constant evolution of formants

Visual inspection of the spectrograms and pitch graphs (I)

- Visible differences in the spectrum richness, large departures from the natural temporal pattern, lack of emphasis (low energy on stressed syllables), and abnormal pitch trajectories are easily identifiable by comparison at the visual inspection.
- The “*s*” fricative is much less energetic in the synthetic voice, the plosive “*t*” is much weaker (almost absent) in the synthetic voice.
- The group “*ci*” is not temporally distinct from the *ne* syllable, “*u*” from *făcut* is too long, and the final syllable “*ta*” is too energetic.

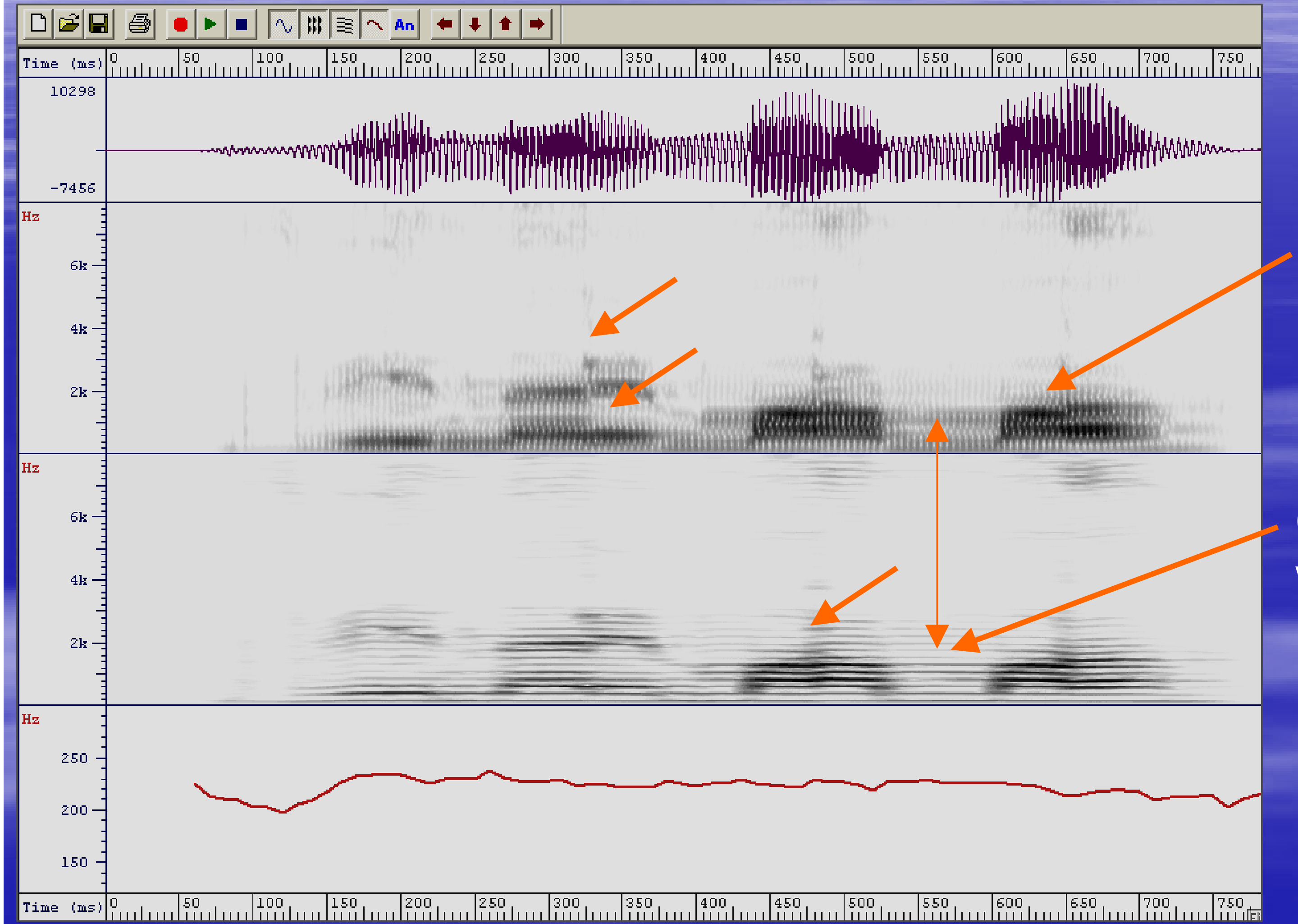
Annotation for human voice using Praat: sentence “Vine mama”



A break of pitch (Fo)

Rich evolution of formants

The evolution of pitch using Wasp; synthetic voice - sentence “Vine mama”



Baum

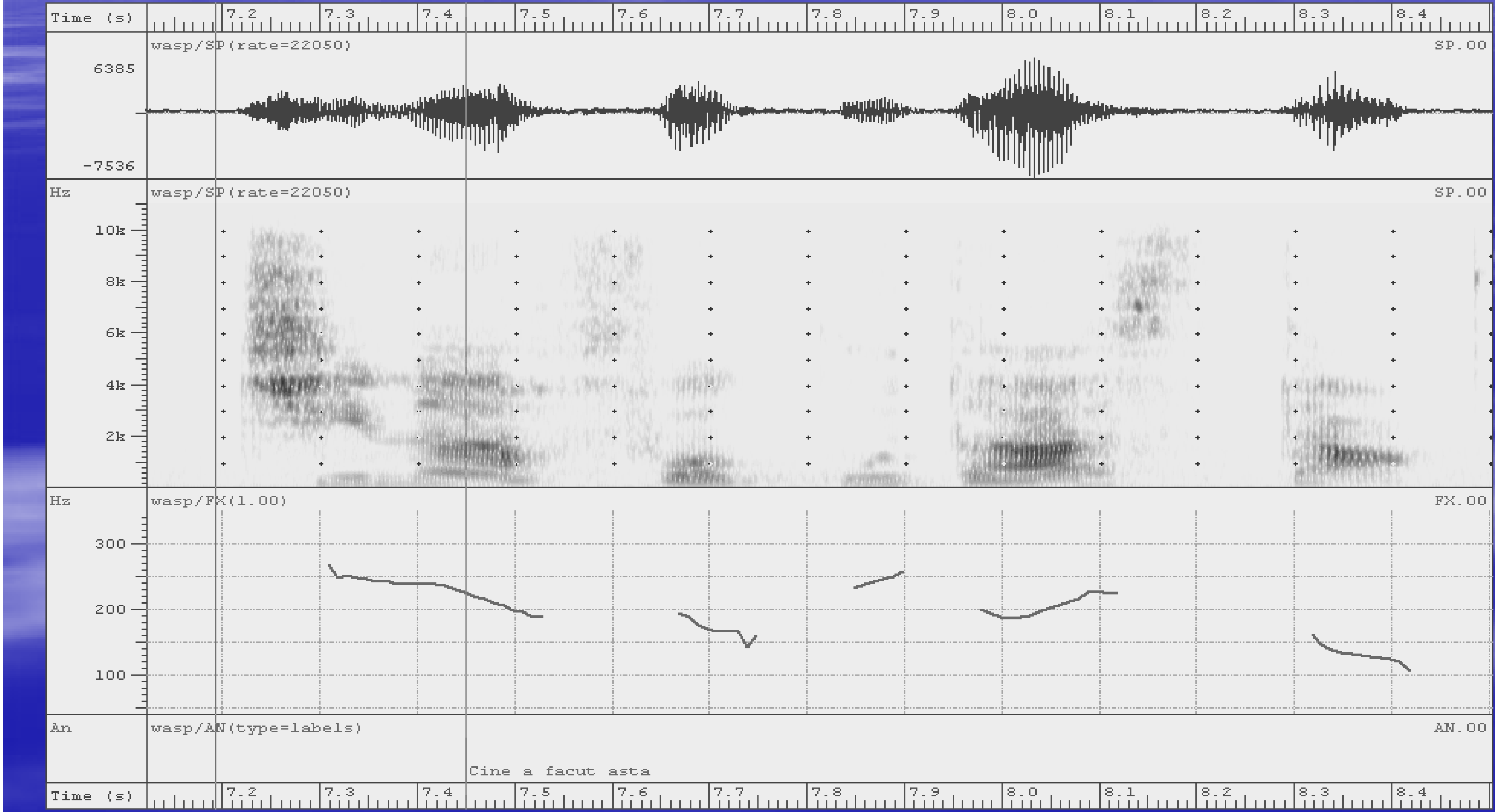
Narrow band

Constant range, without dynamic

Visual inspection of the spectrograms and pitch graphs (III)

- the S1 synthesizer shows that the synthesis overemphasizes the prosody (large variations of the pitch frequency), actor-like, emphatic.
- some errors appearing in the S2 synthesis do not appear in the S1 synthesis: ci is a distinctly spelled sound, “s” is longer and has more energy, and the energy of the final syllable is well proportioned (lower) with respect of the energy of the precedent syllable.
- the S1 synthesis looks better and the spectrogram reveals a rich spectral content of the sounds, with well defined formants.

Visual inspection of the spectrograms and pitch graphs (IV) – synthetic voice – “Cine a făcut asta?”, S1



Analysis of the durations of the sounds-the temporal aspects

- The abnormal long duration of some consonants like t, c, r produces the auditive feeling that the connections between vowels and consonants are malformed. human 🔊 Ivona 🔊 Baum 🔊
- The semi-vocalic consonants (r, n) have durations that are about twice in synthetic speech than for the natural voice, while the plosive consonants (c, t) have a duration more than twice compared to those in natural speech.

The durations of several phones for human and synthesized voices in several sentences

Phonemes		$\overline{d_u}$	d_s	σ_u	$\frac{d_s - \overline{d_u}}{\sigma_u}$	$\frac{\overline{d_u}}{d_s}$
a1	Cine a facut asta	0.075	0.123	0.008	**<-5.5σ**	0.61
c		0.102	0.044	0.015	>3σ	2.32
t1		**0.156**	**0.061**	0.031	>4σ	**5.24**
“**r**” from Aseară		0.065	0.017	0.004	**>12σ**	3.83
i	Ai venit iar la mine	0.081	0.042	0.007	**>5.98σ**	1.94
r		0.075	0.032	**0.026**	>1.62σ	2.31

Results of the numerical quality assessment based on formantic features

$$q_\sigma(F_k/F_0(vowel)) = \begin{cases} 0 & if \quad diff < \sigma[F_k/F_0(vowel,human)], \quad else \\ -1 & if \quad diff < 2\cdot\sigma[F_k/F_0(vowel,human)], \quad else \\ -3 & \end{cases}$$

$$q_\sigma^{total}(vowel) = \underset{k}{Sum}\, q_\sigma(F_k/F_0(vowel))$$

$$Q_{\min} = \min_{vowels}(q_\sigma^{\min}(vowel))$$

$$q_\sigma^{\min}(vowel) = \min_{k} q_\sigma(F_k/F_0(vowel))$$

$$Q_{total} = \frac{1}{N}\cdot \underset{vowels}{Sum}(q_\sigma^{total}(vowel))$$

Comparison of the subjective and the quantitative assessment, for sentence “Vine mama”

	Vowel i	Vowel i	Vowel e	Vowel e	Vowel a1	Vowel a1	Vowel a2	Vowel a2
	S1	S2	S1	S2	S1	S2	S1	S2
F1/F0	0	0	0	0	-3	-3	-3	-3
F2/F0	-3	-3	-3	-3	0	-1	0	-1
F3/F0	0	-1	-3	-3	0	-3	0	-3
L1	RUN	UNN	N	N	UNN	RUN	UNN	UNN
L2	RUN	UNN	SD	SD	RUN	UNN	UNN	UNN

The q^{min}_{σ} and q^{total}_{σ} scores for formant vs. pitch frequencies ratios, for sentence *Vine mama*

Vowels	i	i	e	e	a1	a1	a2	a2
Synthesizer	S1	S2	S1	S2	S1	S2	S1	S2
F1/F0	0	0	0	0	-3	-3	-3	-3
F2/F0	-3	-3	-3	-3	0	-1	0	-1
F3/F0	0	-1	-3	-3	0	-3	0	-3
$q^{total}_{\sigma}(vowel)$	-3	-4	-6	-6	-3	-7	-3	-7

Discussion and Conclusions (I)

- The results obtained show agreement with several characterizations of the synthetic voices performed by human listeners.
- Method based on ratios of values (formants vs. pitch); it is highly sensitive to erroneous measurements in any of these parameters.
- The semi-vocalic consonants (r, n) have durations that are about twice in synthetic speech.

Discussion and Conclusions (II)

- The definitions of quantitative indices offer an easy way to compare the human and the synthesized voices.
- The quality of an utterance should be represented by a vector or, better, by a matrix of individual scores of the phones, moreover of individual features of each phone.

References

1. BAUM Engineering, TTS Online, Voce sintetică românească profesională Ancutza v3.6.16., http://www.baum.ro/index.php?language=ro&pagina=ttsonline
2. Ivona™ synthesizer for Romanian language, http://www.ivona.com/
3. Bruce R. Gerratt, Jody Kreiman, Measuring Vocal Quality with Speech Synthesis. J. Acoust. Soc. Am. Volume 110, Issue 5, pp. 2560-2566
4. Kristin Precoda, Gerald S. Berke, Individual Differences in Voice Quality Perception. Journal of Speech and Hearing Research. Vol. 35 pp. 512-520, June 1992
5. Gobl, C.Bennett, E. Ailbhe Ni Chasaide, Expressive Synthesis: How Crucial is Voice Quality? Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002, pp. 91- 94

Thank you!