

# Preliminary Results of Expressive Speech Synthesis in Polish

Jolanta Bachan<sup>1</sup> & Barbara Surmanowicz<sup>2</sup>

<sup>1</sup>Institute of Linguistics, Adam Mickiewicz University, Poland  
<sup>2</sup>Fachrichtung Elektrotechnik, Wilhelm Büchner Hochschule,  
Germany

*SASR 2008, 8-12 September 2008, Piechowice, Poland*



## young linguists' meeting in poznań

24-26/04/2009

<http://www.ifa.amu.edu.pl/ylmp/>

the world's largest and most influential  
linguistic congresses are held in Poznań  
every four years. The congresses are  
organized by the Institute of Linguistics  
at Adam Mickiewicz University in Poznań.  
The congresses are held in Poznań  
because of the city's rich history  
and culture.

# Objective

- Experiment with expressive speech synthesis

# Objectives

- Experiment with expressive speech synthesis
  - angry, happy, neutral, sad, impatient speech
- Synthesise expressive speech with **neutral** synthetic voices in concatenative synthesis
- Check importance of F0 modelling for expressive speech synthesis, and disregard durations modelling
- Investigate influence on human perception of emoticons when used instead of common word labels

# Available Resources

- Original recordings of male expressive speech
  - 7 different sentences \* 5 expressive states
  - angry, happy, neutral, sad, impatient speech
- MBROLA speech synthesiser
- MBROLA female Polish voice – pl1
- MBROLATOR for MBROLA voice creation
- Automatic Close Copy Speech (ACCS) synthesis software
- Praat
- Diphone extraction software

# Speech Perception Tests

- Test 1: CatNum judgement
  - 7 sentences \* 5 expressive states \* 3 voices = 105

*Listen to the wav file and choose one of the expressive labels which fits the recording best.*

*Mark on the scale the confidence with which you chose the expressive label.  
1 means little confidence, 5 is the highest confidence level.*

- Test 1

# Speech Perception Tests

- Test 2: same / different
  - 1 sentence
    - ACCS of angry, happy, sad, impatient
    - modifications of neutral
      - F0 to angry, happy, sad, impatient
      - durations to angry, happy, sad, impatient
  - AB / BA / AA / BB pairs = 51 pairs
    - AA and BB pairs added as noise to count false alarm rate

*Decide whether the expressive content of the recordings is the same or different, for example, whether both recordings sound happy or not.*

- Test 2

# Speech Perception Tests

- Test 3: emoticons

- 1 sentence \* 5 expressive states \* 3 voices = 15
- compare the results of the same sentences from Text 1 with the results gained from Test 3

*Listen to the wav file and choose one of the emoticons which fits the recording best.*

*Mark on scale the confidence with which you chose the emotional label.*

*1 means little confidence, 5 is the highest confidence level.*

- Test 3

# Material Preparation

- Annotation of speech recordings at phone level
  - for ACCS synthesis
  - for microvoices (diphone databases) creation
- Creating MBROLA microvoices based on male recordings of neutral speech
  - 7 MBROLA microvoices for each neutral sentence
  - MBROLA voice creation process\*
    - automatic extraction of diphones
    - mbrolation with MBROLATOR software

\* We are very grateful to Prof. Dafydd Gibbon and Catharine Oertel for their help and let me use their software

# Synthesis Method: ACCS with MBROLA

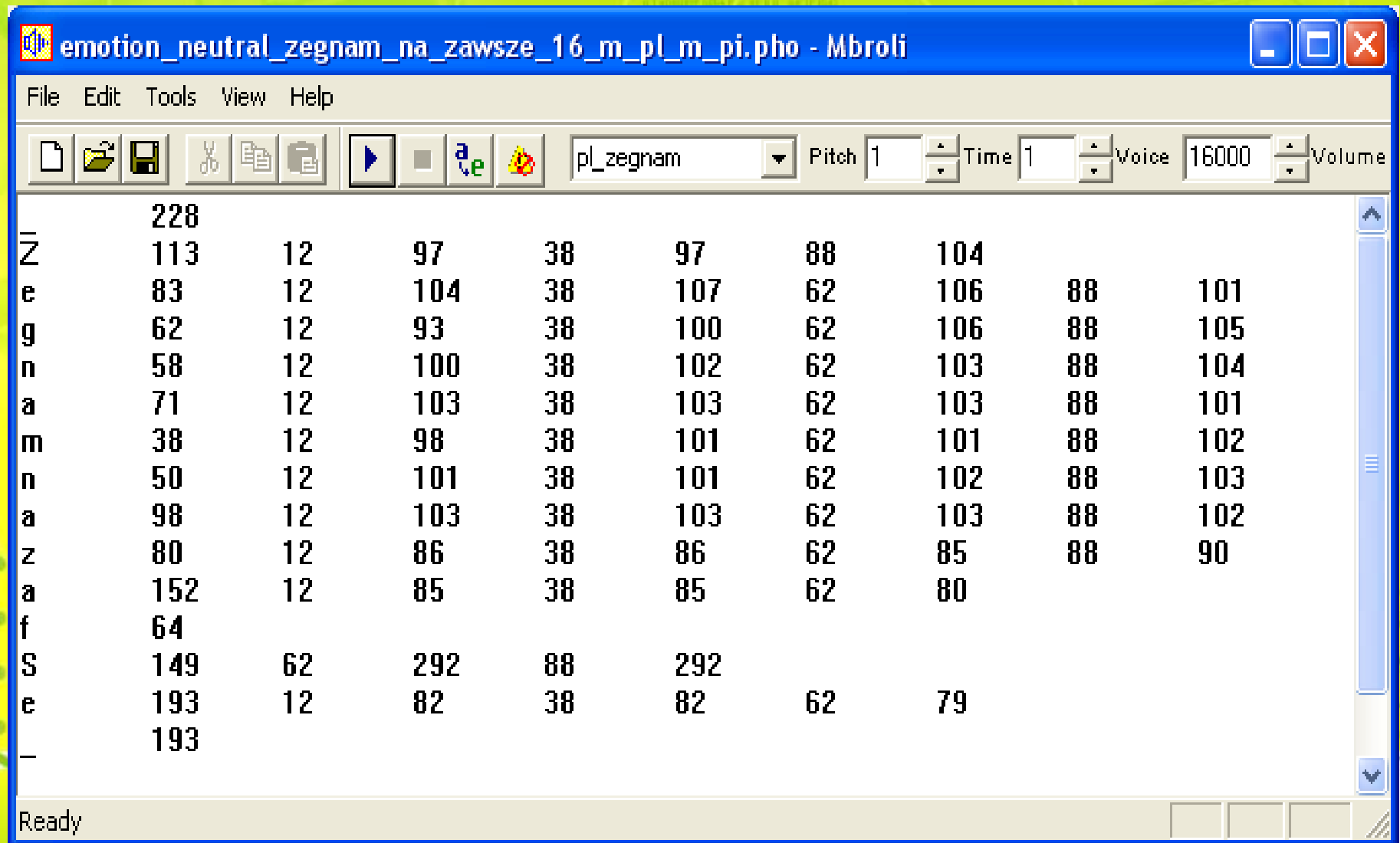
# What is MBROLA?

“MBROLA is a speech synthesiser based on the concatenation of diphones. It takes a list of phonemes as input, together with prosodic information (duration of phonemes and a piecewise linear description of pitch), and produces speech samples of 16 bits resolution (linear), at the sampling frequency of the diphone database used (it is therefore NOT a Text-To-Speech (TTS) synthesizer, since it does not accept raw text as input).” (MBROLA Project website, consulted 2007-05-16)

# Input to MBROLA

- The MBROLA system requires information about
  - Phonemes (transcribed using SAMPA IPA notation)
  - Duration of phonemes (in msec)
  - Pitch pairs
    - pitch height positions within segments (in % of duration)
    - pitch height value (in Hertz)

# MBROLA PHO file table



The screenshot shows the Mbrola software interface with a window titled "emotion\_neutral\_zegnam\_na\_zawsze\_16\_m\_pl\_m\_pi.pho - Mbrola". The interface includes a menu bar (File, Edit, Tools, View, Help), a toolbar with icons for file operations and playback, and a control panel with a dropdown menu set to "pl\_zegnam", and sliders for Pitch (1), Time (1), Voice (16000), and Volume. The main display area contains a table of phonetic data for the text "Zegnana zafse".

	228								
Z	113	12	97	38	97	88	104		
e	83	12	104	38	107	62	106	88	101
g	62	12	93	38	100	62	106	88	105
n	58	12	100	38	102	62	103	88	104
a	71	12	103	38	103	62	103	88	101
m	38	12	98	38	101	62	101	88	102
n	50	12	101	38	101	62	102	88	103
a	98	12	103	38	103	62	103	88	102
z	80	12	86	38	86	62	85	88	90
a	152	12	85	38	85	62	80		
f	64								
S	149	62	292	88	292				
e	193	12	82	38	82	62	79		
-	193								

Ready



# What is Close Copy Speech Synthesis?

The CCS synthesis system produces a sound which “repeats an utterance produced by a human speaker with a synthetic voice, while keeping the original prosody” (Dutoit, 1996).

# ACCS Synthesis Components

- Speech information input
- Speech synthesiser
- Pitch extraction script

# ACCS Synthesis Components

- Speech information input
  - speech recordings
  - time-aligned annotations of speech recordings at phone level in TextGrid format
- Speech synthesiser
- Pitch extraction script

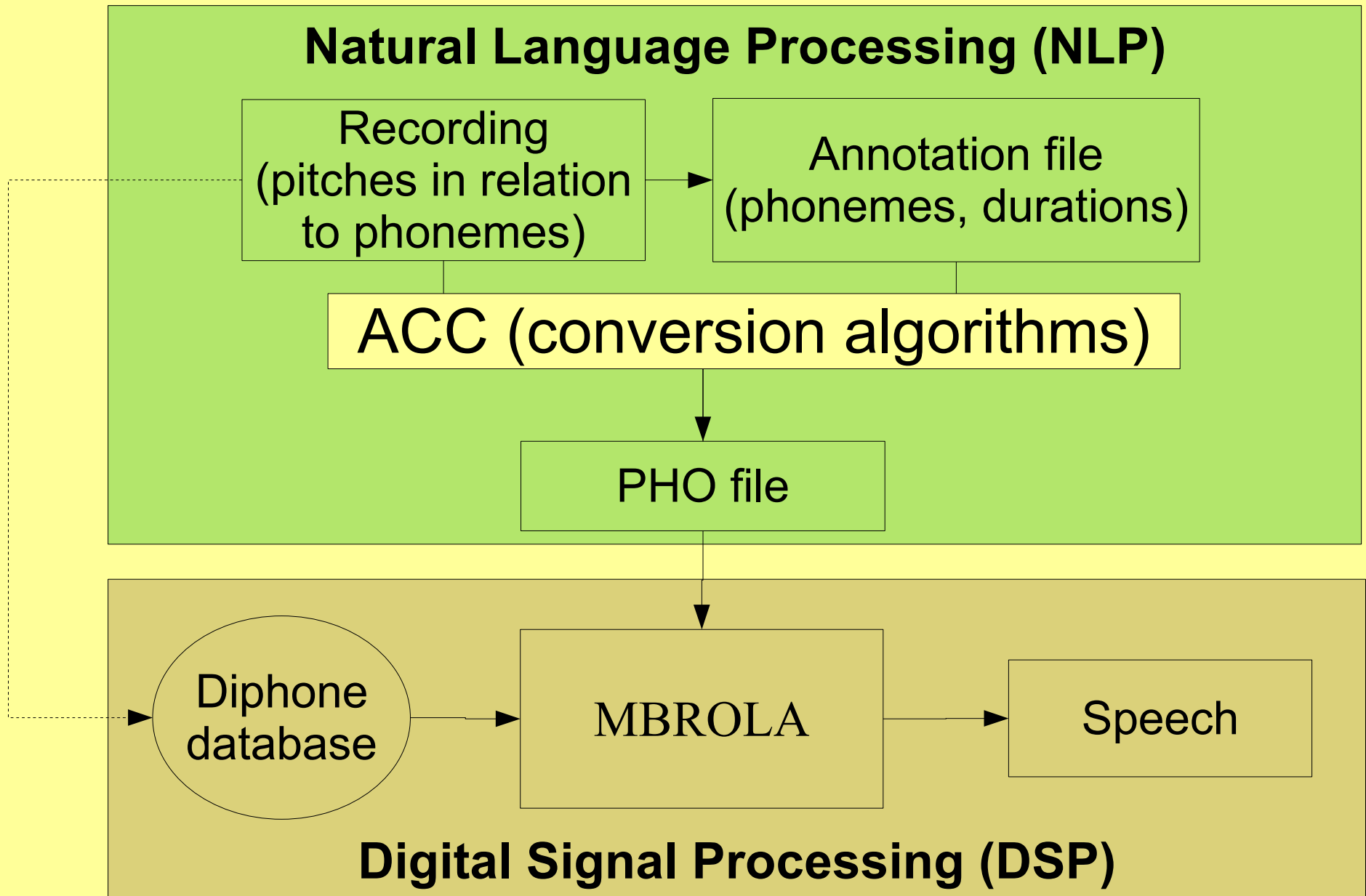
# ACCS Synthesis Components

- Speech information input
  - speech recordings
  - time-aligned annotations of speech recordings at phone level in TextGrid format
- Speech synthesiser
  - diphone database (voice)
  - synthesis engine
- Pitch extraction script

# ACCS Synthesis Components

- Speech information input
  - speech recordings
  - time-aligned annotations of speech recordings at phone level in TextGrid format
- Speech synthesiser
  - diphone database
  - synthesis engine
- Pitch extraction script
  - PROBLEM: mismatch between the voices
    - male recordings, but female MBROLA voice
    - formula: *male pitch value* \* 2

# ACCS Architecture



# Speech Perception Tests Results

- Testees:
  - 6 Polish natives
    - 3 males
    - 3 females
  - 21-28 years old
- The testing session: 30min – 1h 30min






# Results: Test 1 – CatNum judgement

		angry		happy		neutral		sad		impatient	
		Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf
Female	angry	11	3.64	3	2.33	9	2.78	8	3.38	11	3.45
	happy	12	3.58	11	3.45	3	2.67	3	2.67	13	3.46
	neutral	5	3.2	3	2.33	25	3.72	9	3.89	0	-
	sad	2	4	0	-	16	3.38	24	4.08	0	-
	impatient	20	3.35	3	4	0	-	6	3.5	13	3.23
Male	angry	7	3.29	3	4.33	20	3.3	9	3.44	3	3
	happy	7	2.57	9	3.56	14	3	4	3.75	8	2.88
	neutral	4	3.25	0	-	21	4.19	17	4.12	0	-
	sad	1	2	0	-	15	3.47	26	4.15	0	-
	impatient	11	3.45	4	4.25	6	2.67	8	3.63	13	3.54
Original	angry	24	4.17	2	4.5	7	3.43	0	-	9	3.78
	happy	9	3.67	22	3.95	5	3.2	0	-	6	4
	neutral	3	3.33	0	-	27	4.07	10	3.9	0	-
	sad	2	3	1	4	2	3.5	37	4.78	0	-
	impatient	17	4.82	0	-	0	-	0	-	25	4.92

# Results: Test 2 – same / different

	<b>expressive</b>
<b>F0</b>	97.50%
<b>durations</b>	32.50%

# Results: Test 3 – emoticons

		angry				happy				neutral				sad				impatient			
		Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf	Cat	Conf
Female	angry	2	4.5	5	3.2	0	-	1	1	1	2	0	-	2	4	0	-	1	3	0	-
	happy	0	-	0	-	5	3.8	5	3.4	1	4	1	0	0	-	0	-	0	-	0	-
	neutral	1	4	1	3	3	2.33	3	2	2	3	1	4	0	-	1	3	0	-	0	-
	sad	0	-	1	3	0	-	0	-	2	3	4	3	4	4.5	1	4	0	-	0	-
	impatient	5	3.8	3	3	0	-	0	-	0	-	2	2.5	1	5	0	-	0	-	1	1
Male	angry	0	-	1	3	0	-	0	-	6	2.83	5	3.8	0	-	0	-	0	-	0	-
	happy	1	3	1	4	3	4	4	4	2	4	0	-	0	-	0	-	0	-	1	4
	neutral	0	-	0	-	0	-	0	-	3	4.33	2	5	3	4.33	4	3.75	0	-	0	-
	sad	0	-	0	-	0	-	0	-	0	-	5	3.6	6	3.67	1	4	0	-	0	-
	impatient	2	4	2	3.5	0	-	1	4	4	2.75	2	2	0	-	1	4	0	-	0	-
Original	angry	4	3.5	5	4.4	0	-	0	-	2	4.5	0	-	0	-	0	-	0	-	1	4
	happy	2	4.5	3	3.67	4	4.25	3	4.67	0	-	0	-	0	-	0	-	0	-	0	-
	neutral	1	4	0	-	0	-	0	-	2	2	5	3.8	2	4	1	3	0	-	0	-
	sad	0	-	0	-	0	-	0	-	0	-	3	3.33	6	5	3	2.33	0	-	0	-
	impatient	4	4.75	1	5	0	-	0	-	0	-	0	-	0	-	0	-	2	5	5	5

# Summary & Conclusion

- Experiment turned out very successfully
- ACCS and diphone synthesis brought good results
- Expression of the original recordings was correctly recognised
- Expressive F0 and neutral durations made synthetic speech sound expressive
- There is no great influence of emoticons used in test 3 as expressive labels to make the right choice and the confidence is lower than when the word labels were used

Thank you!