Communicative Speech Synthesis with XIMERA: a First Step

Shinsuke Sakai[1,2], *Jinfu Ni*[1,2], *Ranniery Maia*[1,2], *Keiichi Tokuda*[1,3], *Minoru Tsuzaki*[1,4], *Tomoki Toda*[1,5], *Hisashi Kawai*[2,6], *Satoshi Nakamura*[1,2]

[1]NiCT, Japan
[2] ATR-SLC, Japan
[3]Nagoya Institute of Technology, Japan
[4]Kyoto City University of Arts, Japan
[5]Nara Institute of Science and Technology, Japan
[6]KDDI Research and Development Labs, Japan

Introduction

Motivation: high naturalness achieved by concatenative synthesis, but monotonous (Always speaks in the same very articulate way!)

ex. “Taihen moushiwake arimasenga gokouhyounitsuki genzai shinagireto natteorimasu.” (apology spoken rather objectively.)

Goal: synthesizers which can speak in an appropriate style for communicative purposes.

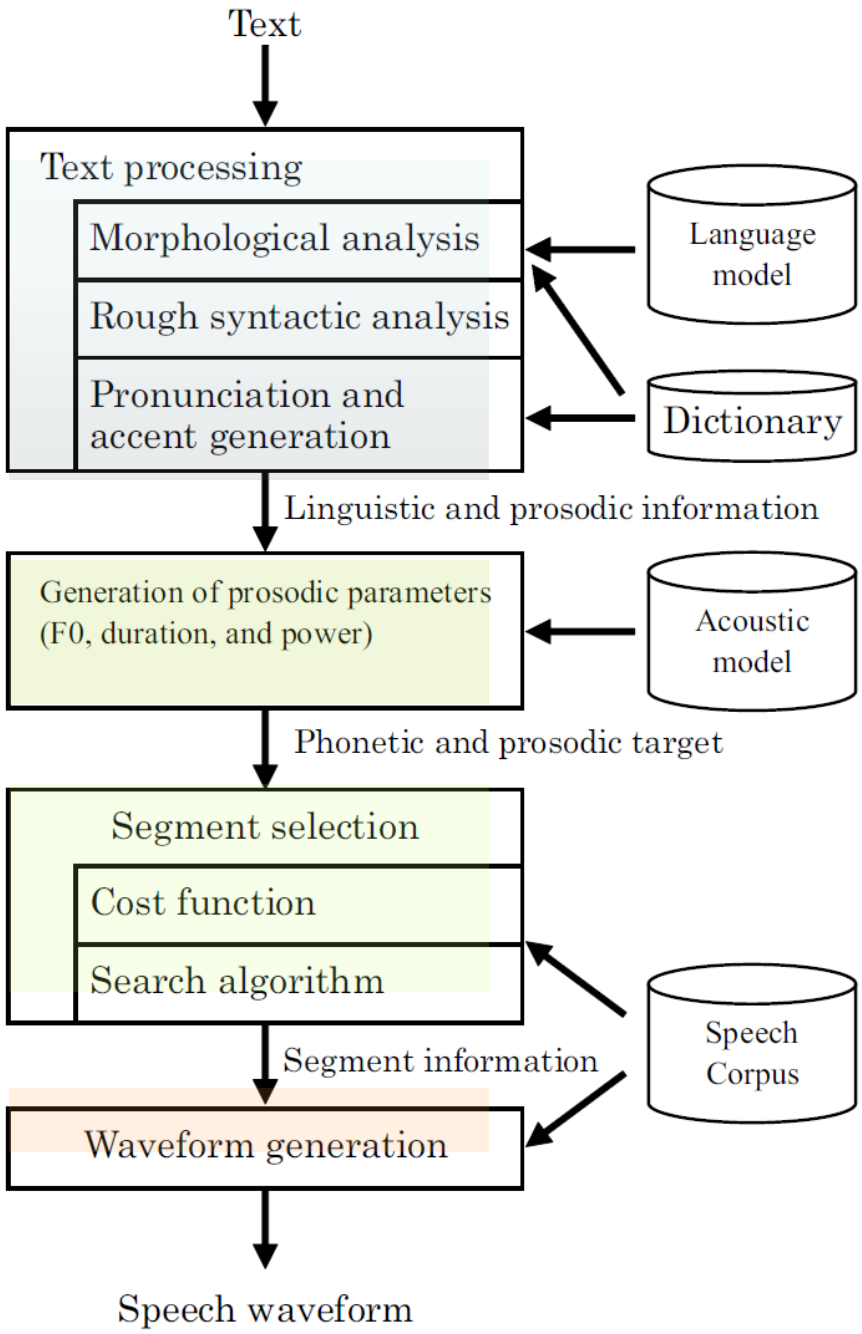
Input extension: ***style tags***

<badnews> There was no room available tonight. </badnews>

Technical approach: Concatenative synthesis (XIMERA) with style-specific target (HMM) and/or style-specific units *(IBM Eide et al. 2003,2004, Pitrelli et al. 2006).*

Overview of XIMERA speech synthesis system

- Large corpora (for Japanese: 110 hours male, 60 hours female)
- HMM target models including prosody.
- Cost functions optimized by perceptual experiments.
- Japanese, Chinese, and English versions.

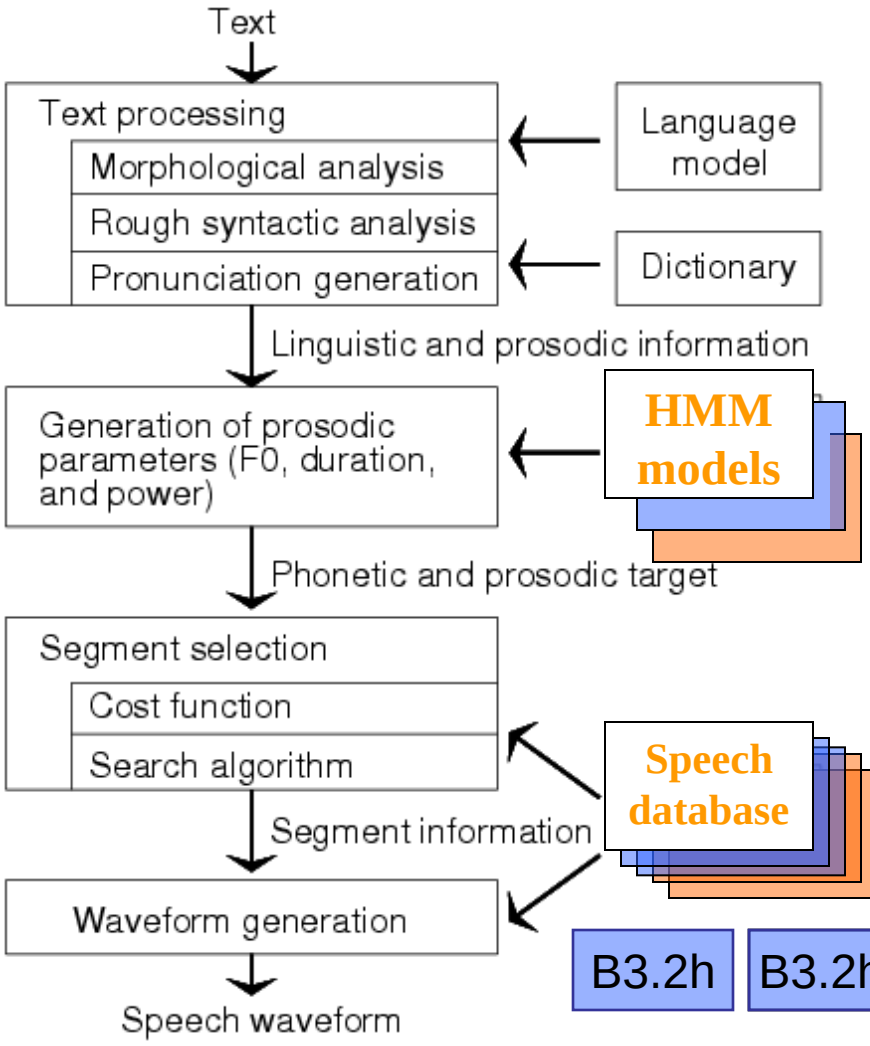


Corpus development

- Speaker: F009 (Japanese female) – 60H neutral DB available.
- Prompt text: extracted subset (corresponding to 2.6 hours of speech) from prompts for neutral DB and modified to have conversational endings.
  - 1/2 Newspaper sentences with conversational utterance-end expressions. (conversion tool with hand correction)
  - 1/4 Phonetically balanced 500 sentence set (ATR503), half as is, half with conversational sentence-ends.
  - 1/4
    - BTEC (basic travel conversation) sentences that can be “news.”
    - Sentences from novels and essays with conversational endings.
- Approximately 3 hours of speech with each of “good news” and “bad news” styles were collected.

Speech corpus	ID	# utterances	# phones	Size
Neutral	N2	1,962	135,142	2.6 h
Good news	G2	1,962	139,551	3.0 h
Bad news	B2	1,962	138,558	3.2 h

Communicative target models and unit databases



XIMERA flowchart

Bad news (2.7h)

Neutral (2.3h)

Good news (2.8h)

B3.2h | B3.2h+N10h | N2.6h | N10h | G3h | G3h+N10h

(B: bad news, N: neutral, G: good news)

Experiment: target models and unit databases

Table 2: *The corpus sizes for training the target HMMs.*

Database style	Size (h)	# utterances	# phones	# feature labels
Neutral	2.3	1,807	118,701	117,638
Good news	2.8	1,852	128,468	126,962
Bad news	2.7	1,726	118,640	117,070

Table 3: *Description of the six databases for use in this test.*

ID	Size	Content
G2	3.0	3 h "good news" speech
G2+N10	13	3 h "good news"+10h neutral speech
N2	2.6	2.6 h neutral speech
N10	10	10 h neutral speech
B2	3.2	3.2 h "bad news" speech
B2+N10	13.2	3.2h "bad news"+10 h neutral speech

Main things we wanted to know

Q1: Are good/bad news styles well observed?

Q2: Do we need style-specific DB, not just style-specific target models?

Q3: Does neutral DB help in naturalness?

8 systems with different combinations of target HMMs and unit DBs.

unit db → *target*↓	G2	G2+ N10	N2	N10	B2	B2+ N10
G	1	2		3		
N			4	5		
B				6	7	8

Experiment: Listening test design

- Test data
  - All sentences carefully designed to be interpreted as good / bad / neutral news.
  - 10 sentences x 8 systems = 80 waveforms.
- Listeners -- 12 native Japanese speakers.
- Experiment I
  - 5-level opinion score on naturalness .. 40 waveforms.
- Experiment II
  - 7-level opinion score on style perception .. 40 waveforms.
    - Ex. -3: sounds like a bad news, -2: pretty sure that it sounds like a bad news, -1: rather sounds like a bad news, 0: no distinction, …

Experiments: results

Table 5: *Percentage of stimuli evaluated as individual levels in Experiment II.*

System ID	–3	–2	–1	0	1	2	3
B–B2	26.7	41.7	30		1.6		
B–B2+N10	26.7	43.3	23.3	6.7			
B–N10	16.7	25	40	11.7	5	1.7	
N–N2			15	63.3	18.3	3.3	
N–N10			16.7	60	18.3	5	
G–G2			3.3	28.3	30	30	6.7
G–G2+N10			3.3	28.3	35	28.3	5
G–N10			0.86	35	46.7	10	

(1)

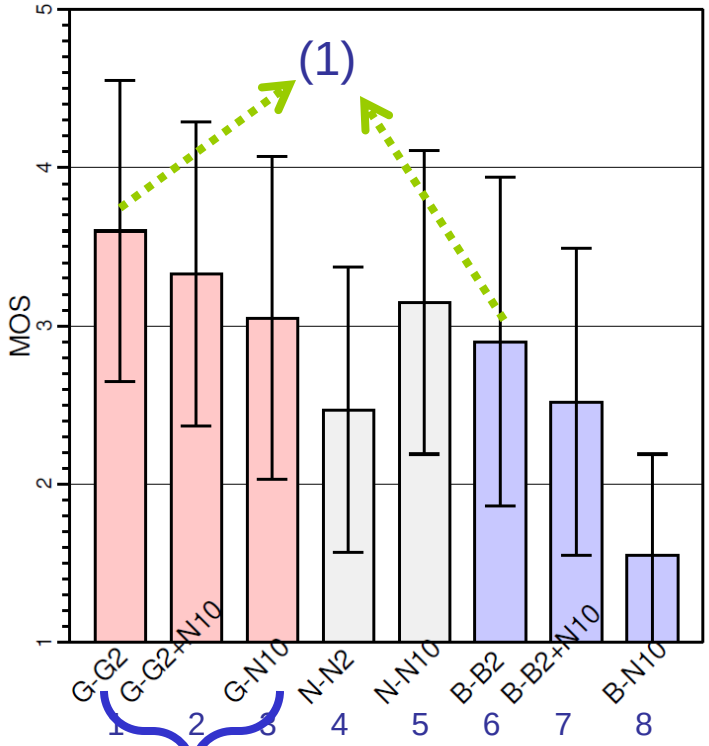


Figure 2: *Mean Opinion Scores and standard deviations for each of the synthesis systems in Experiment I described in the text.*

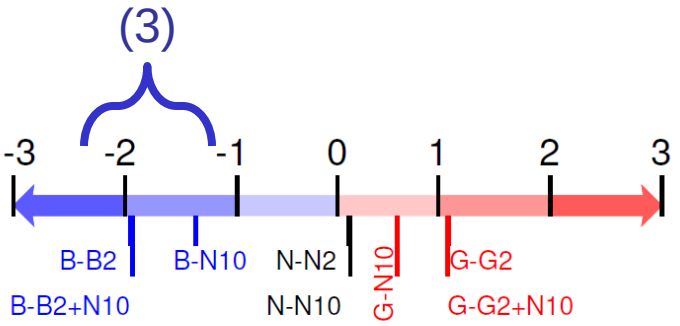


Figure 3: *Mean Opinion Scores for each of the synthesis systems in Experiment II described in the text.*

Observations

(1) Intended style perception well achieved while maintaining a good naturalness.
- "Good news" recognized 66.7%, MOS 3.6 (G-G2)
- "Bad news" recognized 98.4%, MOS 2.9 (B-B2)

(2) "good news" styles sounded more natural to listeners.
- → "good news" more similar to neutral (..?)

(3) Clearer style perception for "bad news".

Experiments: results (cont'd)

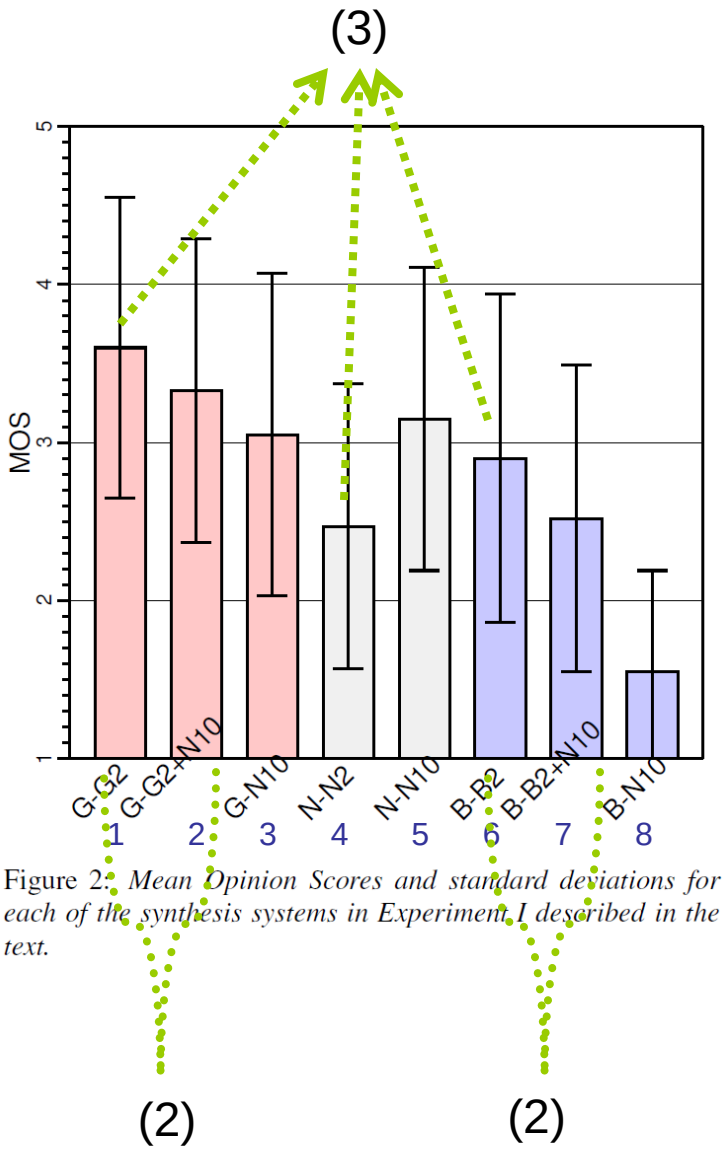


Figure 2: *Mean Opinion Scores and standard deviations for each of the synthesis systems in Experiment I described in the text.*

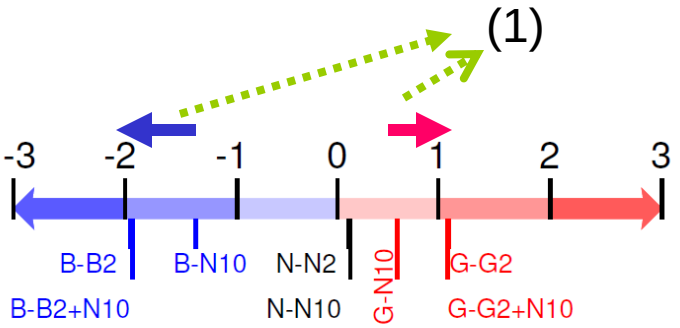


Figure 3: *Mean Opinion Scores for each of the synthesis systems in Experiment II described in the text.*

Other observations

(1) Target alone is not enough. Unit DB for the specific style makes difference.

(2) Addition of neutral data doesn't improve naturalness (a little degradation instead).

(3) Speech with good/bad news styles sounded more natural if developed with the same amount of data.

Experiments: F0-related observations

Table 7: *F0 ranges for the three styles in a speaker.*

Style	Mean	Standard deviation	F0 range
Bad news	161 Hz	33 Hz	(105 Hz, 242 Hz)
Neutral	249 Hz	56 Hz	(131 Hz, 365 Hz)
Good news	283 Hz	61 Hz	(167 Hz, 407 Hz)

- Natural F0 for:
  - –"bad news" speech:
    - F0 mean is low
    - dynamic range is narrower
  - –"good news" speech:
    - F0 mean a little higher.
    - Dynamic range little wider.

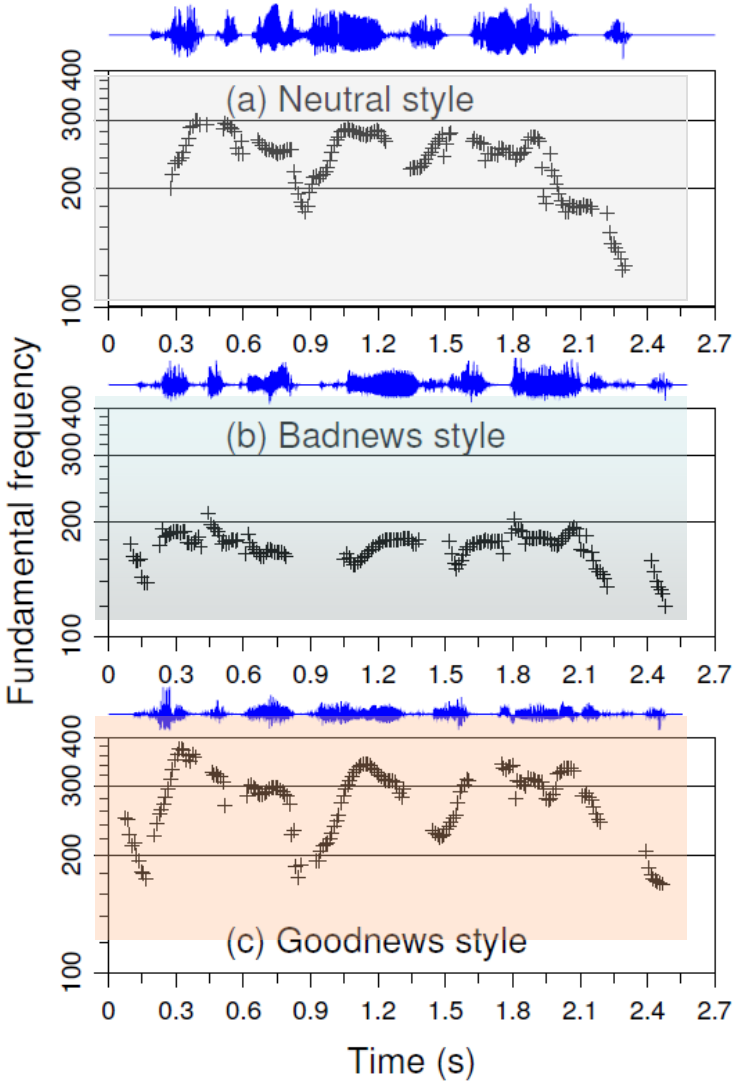


Figure 4: *Examples of F0 contours for a Japanese sentence produced by a native in (a) neutral styles, (b) "bad news" styles, and (c) "good news" styles.*

Conclusion

- Initial attempt at communicative speech synthesis with “good news” and “bad news” styles using 3 hours of each style-specific corpora.
- Intended style perception well achieved while maintaining a good naturalness. “Good news” recognized at 66.7% with MOS 3.6 (G-G2). “Bad news” recognized at 98.4% with MOS 2.9 (B-B2).
- Not only target models but also unit databases with specific styles were effective in synthesizing speech in the intended corresponding styles.
- Plan to investigate contributions from each of spectral, F0, and duration features separately, instead of the models themselves.

appendices

(appendix) test sentences

・neutral, good news, bad news いずれの解釈も可能な文セットを用意する。

input01 ご主人の体重は先月から5キログラム増えています。
input02 気がついたら、うちの庭にタンポポの花が咲いていました。
input03 さきほど述べた点が第一の要因だと考えられます。
input04 先月はまだ130万円でしたが、今月は170万円になっています。
input05 そのOBは、退職と同時に海洋土木大手企業で勤務していたことがわかりました。
input06 トヨタはグループ各社のトップに出身者を送り、結束強化を図ってきたそうです。
input07 東亜建設は、同社に技術指導の名目で100万円の委託料を支払いました。
input08 TBSは、この内容を楽天に文書で通告しました。
input09 米国では、新聞業界に再編の動きがあります。
input10 71年6月に設立された同社には約30人が社員として在籍していました。

5段階評価

- どのくらい自然に聞こえますか？ (人間の声と区別がつかない音の点数を5とします。)

1　2　3　4　5

7段階評価

- 良い知らせに聞こえますか。それとも、悪い知らせに聞こえますか?

-3 -2 -1 0 1 2 3

-3: 悪い知らせに聞こえる。
-2: ほぼ間違いなく悪い知らせに聞こえる。
-1: どちらかというと悪い知らせに聞こえる。
0: どちらともいえない。
1: どちらかというと良い知らせに聞こえる。
2: ほぼ間違いなく良い知らせに聞こえる。
3: 良い知らせに聞こえる。