Intelligent System Design
~ Speech Generation ~

7th Lecture
Nov. 21st, 2011

Tomoki Toda
Augmented Human Communication Laboratory
Speech Communication

- Speech is a very effective way for us to communicate with each other.

**Speaker**

- Concept
- Semantic encoding
  - Linguistic/paralinguistic information
  - Signal encoding
  - **Speech**

**Listener**

- Concept
- Semantic decoding
  - Linguistic/paralinguistic information
  - Signal decoding
  - Others such as gesture, facial expression, and so on
Effectiveness of Speech

- Speech is a very useful communication medium because speech can convey various types of information simultaneously.

Listen to this speech signal:

What information is conveyed?

What kinds of information can be conveyed with this one-dimensional feature?
• In speech communication, conceptual information is effectively represented with spoken language.

**Conceptual information**

**Spoken language**

- Words
  - Grammar
  - Syntax
  - Sentences

“I am hungry. Let’s go to a Japanese restaurant.”
In speech communication, not only linguistic information but also paralinguistic information is conveyed.
- Emotions, attitudes, intentions, ...

Even if using the same sentence, different meanings can be conveyed, *e.g.*, 
- I didn’t take the test yesterday. (Somebody else did.)
- I *didn’t* take the test yesterday. (I did not take it.)
- I didn’t *take* the test yesterday. (I did something else with it.)
- I didn’t take *the* test yesterday. (I took a different one.)
- I didn’t take the *test* yesterday. (I took something else.)
- I didn’t take the test *yesterday*. (I took it some other day.)
Non-Linguistic Information

• In speech communication, not only linguistic information and paralinguistic information but also nonlinguistic information is conveyed.
  – Gender, age, state of health, speaker personality, …

• Nonlinguistic information is not intentionally conveyed but it is helpful for our speech communication.

My son is speaking.
He seems to have a cold…
Speech Generation Process

- Speech generation is required in some applications, such as spoken dialog systems.

Dialog management

Concept

What to say?

Semantic representation

CONFIRM ( TO ( CITY ( Dallas ) )
ON ( DATE ( Monday ) ) )

Nonlinguistic information

Female voice

Paralinguistic information

Emphasis: “Monday”

Natural language generation (NLG)

How to say?

Text

“You want to go to Dallas on Monday, don’t you?”

Text-to-Speech (TTS)

How to utter?

Speech waveform
An Example in Spoken Dialog System

Speech recognition hypotheses

Demo movie of Spoken Dialog Systems group at Cambridge University
http://mi.eng.cam.ac.uk/research/dialogue/

TTS input (NLG output)

System beliefs about possible user goals

Spoken language understanding hypotheses

System dialogue act outputs (semantic representation)
**Natural Language Generation (NLG)**

- NLG is a technique to convert semantic representation (e.g., tree) into text (i.e., word sequences).

**Semantic representation** \(\xrightarrow{\text{How to say?}}\) **Text**

- Rule-based approach
- Statistical approach

**Discontinuous to discontinuous mapping**

CONFIRM ( TO ( CITY ( Dallas ) )
ON ( DATE ( Monday ) ) )

Hierarchical structure
Mechanism of Speech Production

- **Articulation**: Adding various sounds by adjusting the vocal tract shape
- **Excitation**: Generating air vibration by the opening and closing of vocal folds

Speech signal → Convolution
Mechanism of Excitation Generation

- Excitation sounds are produced as the air vibration generated with the opening-closing of vocal folds.
  - Normal condition
  - Speaking condition
  - Playing in slow motion

- Higher pitched voices are generated by making the opening-closing timing faster.
- Lower pitched voices are generated by making the opening-closing timing slower

- These photos and movie come from http://www.osaka-med.ac.jp/deps/oto2/onsei_2.html
Glottal Excitation Waveform

- **Waveform**
  - Zoom in

- **Volume of voice**
  - Pitch of voice
    - Higher pitch → Shorter period
    - Lower pitch → Longer period
Mechanism of Articulation

• Various sounds are produced according to the resonance characteristics determined by the vocal tract shape.

Several frequency components are enhanced by the resonance.

Spectral envelope: Resonance characteristics

![Diagram showing frequency components and spectral envelope.]
Excitation and Speech Waveforms

Excitation waveform

Convolution of impulse response of spectral envelopes

Speech waveform
Speech Features

**Vocal tract feature**
- Capturing phoneme sounds and voice quality
  - Spectral envelope

**Excitation feature**
- Controlling prosodic information such as intonation
  - Fundamental frequency ($F_0$)
Segmental and Prosodic Features

• **Segmental feature**
  – Mainly depending on individual phonemes
  – Examples: spectral envelope, …

• **Prosodic feature (or supra-segmental feature)**
  – Mainly depending on words, phrases, and sentences rather than individual phonemes
  – Prosody, intonation, rhythm, …
  – Examples: $F_0$ pattern, phoneme duration, …
An Example of Speech Features

- Speech waveform is effectively represented by a few speech features.

Speech waveform

$F_0$ pattern

Spectral envelope

Phoneme duration

「あらゆる現実を・・・」

Spectral envelope

Phoneme duration

$F_0$ pattern

Speech waveform
Text-to-Speech (TTS)

- TTS is one of the speech synthesis techniques.
- An arbitrary input text is converted to a speech waveform.

How to utter?
- Rule-based approach
- Statistical approach

Discontinuous to continuous mapping
Samples of Synthetic Speech (1)  

[Klatt, 1987]

Development of speech synthesizers

- VODER, Homer Dudley, 1939.
- PAT (Parametric Artificial Talker) parallel formant synthesizer designed by Walter Lawrence, 1953.
- OVE (Orator Verbis Electris) cascade formant synthesizer, Gunnar Fant, 1953.
- Comparison of synthesis and a natural recording, automatic analysis-resynthesis using multipulse linear prediction, Bishnu Atal, 1982.

Segmental synthesis by rule


Synthesis by rule of sentences and sentence prosody

Samples of Synthetic Speech (2) [Klatt, 1987]

**Fully automatic text-to-speech conversion**
- The first full TTS system, Noriko Umeda et al., 1968.
- The Klattalk system, Dennis Klatt, 1983.
- AT&T Bell Laboratories TTS system, 1985.
- Several of the DECtalk voices: Perfect Paul, Beautiful Betty, Huge Harry, Kit the Kid, and Whispering Wendy.

A statistical approach (called a corpus-based approach) has been proposed [Sagisaka, 1988] and it has enabled significant progress in speech synthesis.

**Products**
- Natural voices of AT&T
- RealSpeak of Nuance

**Free software**
- Festival from CMU
- HMM-based speech synthesis system (HTS) from NITECH
To achieve natural and smooth man-machine communication, what kinds of functions need to be implemented for NLG and TTS?

– Which functions have already been implemented?
– Which ones have not been implemented yet?
NLG

Semantic representation → Text
Template-Based NLG

- Word or phrases in carrier sentences are changed according to semantic representations.
  - Merits: easy to implement
  - Demerits: hard to manage if a lot of semantic representations are used

Semantic representation

“area の date の info-type は weather です．”

“奈良北部の明日の天気は晴れです．”
Statistical Approaches to NLG

- Statistical approaches are capable of developing trainable NLG systems that can be adapted to a new domain.
- Ranking-based approach [Stent et al., 2004]
  - Large varieties of sentences are acceptable for a single semantic representation.
  - Possible sentences are generated from a semantic representation (e.g., using hand-crafted rules).
  - Parameters of ranking process is automatically trained.

```
Semantic representation

Sentence 1  Sentence 2  ...  Sentence N

Ranking

Output sentence
```

Trainable ranking model (such as RankBoost [Schapire, 1999], [Freund et al., 2003])
Finite-State Tagger (FST) Model

[Levin and Pieraccini, 1995]

- Relationship between word sequence and semantic label sequence is probabilistically modeled.
  - Word sequence: $W = \{w_1, \ldots, w_T\}$
  - Semantic label sequence: $C = \{c_1, \ldots, c_T\}$

Probability of word sequence:

$$P(W) = \sum_{\text{all } C} P(C)P(W | C)$$

$n^{th}$ order Markov assumption

$$= \sum_{\text{all } C} \prod_{t=1}^{T} P(c_t | c_{t-1}, \ldots, c_{t-n+1})P(w_t | w_{t-1}, \ldots, w_{t-n+1}, c_t)$$
• **Semantic vector** is used to model a hierarchical representation.

Semantic representation

```
RETURN ( TOLOC ( CITY ( Dallas ) )
ON ( DATE ( Thursday ) ) )
```

Semantic vector sequence
Definition of HVS

Word sequence: \( W = \{w_1, \cdots, w_T\} \)

Semantic vector sequence: \( C = \{c_1, \cdots, c_T\} \)

Semantic concept labels: \( c_t = c_t[1\cdots D_t] \)

Probability of word sequence:

\[
P(W) = \sum_{\text{all } C} P(C)P(W \mid C)
\]

1\textsuperscript{st} order Markov assumption

\[
= \sum_{\text{all } C} \prod_{t=1}^{T} P(c_t \mid c_{t-1})P(w_t \mid c_t)
\]

Conditionally independence assumption

Modeled with stack operation (pop & push)

\[
= \sum_{\text{all } C} \prod_{t=1}^{T} P(n_t \mid c_{t-1})P(c_t[1] \mid c_{t-1}[(n_t + 1)\cdots D_{t-1}])P(w_t \mid c_t)
\]

\( n_t \) labels are popped off from the stack \( c_{t-1} = c_{t-1}[1\cdots D_{t-1}] \)

The label \( c_t[1] \) is pushed on the stack \( c_{t-1}[(n_t + 1)\cdots D_{t-1}] \)
Example of Stack Operation

- Stack operation from $t = 7$ to $t = 8$, i.e., “… Dallas on …”
  - Two conceptual labels are popped off.
  - One conceptual label is pushed on.

$\begin{align*}
  n_8 &= 2 \\
  c_7 &= c_7[1\cdots4] \\
  c_8 &= c_8[1\cdots3]
\end{align*}$

Joint probability at $t = 8$:

$P(n_t \mid c_{t-1})P(c_t[1] \mid c_{t-1}[(n_t + 1)\cdots D_{t-1}]})P(w_t \mid c_t)$
Decoding and Generation

- HVS is used as both decoding and generation processes.
  - Decoding process (word sequence into semantic vector sequence):
    \[
    \hat{C} = \arg \max_c P(C \mid W) \\
    = \arg \max_c P(C)P(W \mid C)
    \]
  - Generation process (semantic vector set into word sequence):
    \[
    \hat{W} = \arg \max_w P(W \mid C^0) \\
    = \arg \max_w \sum_{c \in C_{set}} P(C)P(W \mid C)
    \]
  - Approximation:
    \[
    \hat{C} = \arg \max_c P(C) \\
    \hat{W} = \arg \max_w P(W \mid \hat{C})
    \]

\(C_{set}\): All possible semantic vector sequence
Phrase-Based Generation

- A phrase sequence is generated from a semantic vector sequence.

\[
\hat{W} = \hat{R} = \arg \max_R \prod_{t=1}^{T} P(r_t \mid r_{t-1}, h_t, l_{t-1}, l_t, l_{t+1}, c_{t-1}, c_t, c_{t+1})
\]

Phrase sequence

\[\begin{align*}
&\text{Realization phrase } r \\
&\text{Stack tail } l \\
&\text{Stack head } h \\
&\text{Semantic vector } c
\end{align*}\]

\[\begin{array}{ccc}
t = 1 & t = 2 & t = 3 \\
\text{Deterministic} & \text{Observed}
\end{array}\]

[Francois et al., 2010]
Note that I will give a lecture about details of TTS in Speech Processing 2 in the 4th quarter (Dec. 5 and 7).
Two Main Approaches to TTS

• Rule-based approach (—1990)
  – Hand-crafted rules to generate speech features are developed by individual researchers.
    • Analyzing speech samples, implementing generation rules, synthesizing speech samples, evaluating …
    • Hard to extract reasonable and generic rules…

• Corpus-based approach (1990—)
  – Using a large amount of speech data, the process to generate speech features are automatically trained.
    • Describing speech synthesis process mathematically, implementing machine learning methods, …
    • Easy to develop systems using generic algorithms
Framework of TTS

Text

“小さな鰻屋に，熱気のようなものがみなぎる．”

Text analysis

Phonemes, accentual information, …

Prosody generation

$F_0$ pattern, phoneme duration, …

Segment generation

Waveform/spectral segments

Waveform generation

Speech database

Synthetic speech signal
Text Analysis

• Text normalization
  – Removing redundant parts, such as header, ...

• Morphological analysis
  – Sentence into morphemes
  – Reading and accent

• Dependency parsing
  – Break estimation

• Phoneme generation
  – Generating phoneme sequences

• Accentual phrase processing
  – Determining accentual phrases and their accent types
Japanese sentence is segmented into a row of morphemes.
Not only part-of-speech tagging but also word segmentation is required for Japanese.
Various statistical methods, such as N-gram, HMM, MEMM, CRF, and so on, have been developed.

<table>
<thead>
<tr>
<th>本社</th>
<th>で</th>
<th>生産</th>
<th>や</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>Case particle</td>
<td>Noun</td>
<td>Coordinating particle</td>
</tr>
<tr>
<td>本社</td>
<td>で</td>
<td>生産</td>
<td>や</td>
</tr>
<tr>
<td>Prefix</td>
<td>Suffix</td>
<td>Noun</td>
<td>Prefix</td>
</tr>
<tr>
<td>ほん</td>
<td>しゃ</td>
<td>で</td>
<td>なま</td>
</tr>
<tr>
<td>Noun</td>
<td>Noun</td>
<td>Verb</td>
<td>Noun</td>
</tr>
</tbody>
</table>
• In the following three sentences, phoneme sequences are equivalent to each other but their linguistic information is completely different.
  
  - 「箸を見る。」 ハシヲミル
  - 「橋を見る。」 ハシヲミル
  - 「端を見る。」 ハシヲミル
  
  * These three words, 箸, 橋, and 端, have different accent types.

• Japanese accent is controlled by a pitch ($F_0$) pattern.
  - Accent is given by rapidly decreasing $F_0$ (i.e., high to low).

• Accentuation rules are usually used to estimate an accent type of an accentual phrase consisting of multiple words.
**Prosody Generation**

- Prosodic features, such as $F_0$ pattern and phoneme duration are estimated from linguistic features extracted in text analysis.
  - **Estimation of phoneme duration**
    - Linguistic features, *e.g.*, - current phoneme, - phoneme environment, - part-of-speech, - speaking rate, …
  - **Estimation of $F_0$ pattern**
    - Linguistic features, *e.g.*, - accentual phrase boundary, - accent type, - phrase boundary, - speaking rate, …

Prediction model, *e.g.*, - linear regression, - decision tree, - neural network, …

$f_0$ pattern

Trained from speech corpus

Phoneme durations
Segment Generation

- Speech feature sequences in a speech corpus are segmented to basic units and they are stored.
- In synthesis, the stored unit segments are concatenated to generate segment features.
  - Various types of basic units are used, *e.g.*,  
    - Phoneme
    - Diphone
    - CV (C: consonant, V: vowel)
    - VCV
    - CVC
    - CV*
    - Non-uniform unit
    - Context-dependent units
    - Prosodic-context-dependent units
Segment Selection

The optimum segment sequence is dynamically selected according to input text, and the selected segments are concatenated.

Target (linguistic/prosodic features)

Segment candidates in speech corpus

All segments of phoneme “e” in speech corpus

Selected segment sequence

Concatenated segments

[Iwahashi et al., 1992] [Hunt and Black, 1996]
Waveform Generation

- Speech waveform is generated by concatenating waveform segments or synthesized from concatenated speech features (spectral envelope) and the generated $F_0$ pattern.
  - Waveform concatenation

  ![Waveform Concatenation Diagram]

  - Speech feature concatenation and vocoder (source-filter model)

  ![Speech Feature Concatenation Diagram]
HMM-Based Speech Synthesis

[Tokuda et al., 2000] [Tokuda et al., 2002]

- Speech feature sequences are modeled by HMMs in training.
- In synthesis, speech features are directly generated from HMMs, and then, synthetic speech waveform is generated.
Expressive Speech Synthesis

- Due to a complete data-driven approach, paralinguistic information and nonlinguistic information in speech corpus are also modeled in the HMM-based speech synthesis system.

- Synthetic speech is easily transformed by modifying HMM parameters in HMM-based speech synthesis.
  - Capable of mimicking, mixing, manipulating voices

- Emphasis is also easily implemented for HMM-based speech synthesis.
Mimicking Voice (Adaptation)

- Model adaptation techniques are available to develop TTS with desired voices using a small amount of speech data.
  - Adaptation to different speakers
    
    | Initial model (avg.) | Male 1 | Male 2 | Male 3 | Female |
    |----------------------|--------|--------|--------|--------|
    | Target vocoded voice |        |        |        |        |
    | Speaker-dependent model |        |        |        |        |
    | Adapted model        |        |        |        |        |

  - Adaptation to different speaking styles
    
    | Style-dependent model | Joyful | Rough | Sad |
    |------------------------|--------|-------|-----|
    |                        |        |       |     |
    | Adapted model          |        |       |     |

Samples from HP of Dr. Yamagishi of University of Edinburgh
Voice morphing is easily achieved by interpolating parameters of different HMMs.

- Morphing between different speaking styles
  - Gradually morphed from sad to joyful:
  - Gradually morphed from joyful to sad:

Samples from HP of Dr. Yamagishi of University of Edinburgh
Manipulating Voice (Regression)

- Speaking style is manually controlled by performing linear regression between HMM parameters of different speaking styles and style control vectors to be manipulated.

Samples from HP of Dr. Yamagishi of University of Edinburgh
References