Prosody control in HMM-based speech synthesis
Sathish Pammi, DFKI GmbH
Saarbrücken, Germany

This report describes our work on controlling prosody in Hidden Markov Model (HMM) based speech synthesis using MARY TTS. The MARY TTS platform (Schröder et al, 2008; Schröder et al, 2003) is an open-source, modular architecture for building text-to-speech systems, including unit-selection and HMM based synthesis technologies.

HMM-based speech synthesis

Traditionally, the applications that require control over prosody were using MBROLA diphone synthetic voices though the voices are unnatural. Nowadays HMM-based voices, which can support prosody modification, are reaching high quality synthetic speech.

In HMM-based speech synthesis, trained statistical models (context-dependent HMMs) are used to predict duration and generate parameters like mel-cepstral coefficients, log F0 values, and bandpass voicing strengths using the maximum likelihood parameter generation algorithm including global variance (Toda et al, 2007). In the later stages, F0 parameters, bandpass voicing strengths, and the five bandpass filters are used to generate a mixed excitation signal. Finally, speech is synthesized from the mel-cepstral coefficients and the mixed excitation signal using the MLSA filter.

Controlling prosody with MARYXML

The recent development in MARY framework support reliable prosody modification using 'prosody' element (See MARYXML 1). The 'prosody' element is well described in W3C Speech Synthesis Markup Language (SSML) recommendations. The different attributes in 'prosody' element like 'rate', 'pitch' and 'contour' are used as specifications to modify predicted phone durations and pitch contour before passing them to the HMM synthesizer. Once such modifications are done according to given specifications, they are realized as normal with HMM-based synthesis strategies (Tokuda et al, 2002; Toda et al, 2007).

<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns=" http://mary.dfki.de/2002/MaryXML"
xml:lang="en-US">
<p>
<prosody rate="fast"
pitch="+10%"
contour="(10%,low) (80%,+10%) (100%,+5st)"
>
Welcome to the world of speech synthesis!
</prosody>
</p>
</maryxml>

MARYXML 1: An example which supports prosody specifications
MARYXML Examples

1. Adjusting speech rate

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="en-US">
  <p>
    <prosody rate="200%">
      I am now speaking twice as fast.
    </prosody>
  </p>
</maryxml>
```

MARYXML 2: Example of speech rate modification

2. Adjusting pitch level

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="en-US">
  <p>
    <prosody pitch="-5st">
      This is five semitones lower pitch.
    </prosody>
  </p>
</maryxml>
```

MARYXML 3: Example of pitch level modification

3. Shaping-up pitch contour

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="en-US">
  <p>
    <prosody contour="(2%,+5%)(85%,+5%)(100%,+50%)">
      That is a red ball.
    </prosody>
  </p>
</maryxml>
```

MARYXML 4: Example of pitch contour shaping for high intonational rise in the final part of the utterance

For more information on the MaryXML `<prosody>` markup which can now be applied to all types of MARY voices, see [http://mary.opendfki.de/wiki/ProsodySpecificationSupport](http://mary.opendfki.de/wiki/ProsodySpecificationSupport)
Intonation control using ToBI accents

As described above, MARY framework is supporting prosody specification to control intonation of HMM voices. To make another possibility to control intonation using ToBI accents too, we implemented a conversion mechanism that uses a ToBI accents to intonation contour shapes look-up. The TOBI accents are defined as pitch contour specifications in a look-up file shown in Table 1 (Contour shapes are referred from Huang et al 2001).

<table>
<thead>
<tr>
<th># ToBI accents to f0 contour lookup</th>
</tr>
</thead>
<tbody>
<tr>
<td># format: ToBI accent key</td>
</tr>
<tr>
<td>H*</td>
</tr>
<tr>
<td>L*</td>
</tr>
<tr>
<td>L*+H</td>
</tr>
<tr>
<td>L*+!H</td>
</tr>
<tr>
<td>L+H*</td>
</tr>
<tr>
<td>!H*</td>
</tr>
</tbody>
</table>

Table 1: Look-up table for ToBI accents to intonation contour shapes

According to Table 1, currently, MARY supports intonation control for H*, L*, L*+H, L*+!H, L+H*, !H* ToBI accents. For example, MARYXML 5 generate 'That ball is a red ball' utterance with 'H*' kind of emphasis on 'red'.

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="en_US">
<t accent="none">That</t>
<t accent="H*">red</t>
<t accent="none"> ball</t>
<boundary breakindex="5" tone="L-L%" />
</maryxml>
```

MARYXML 5: An example that support intonation control using ToBI accents

**Note 1:** The conversion of ToBI to pitch contour shapes are dependent on a boolean property called "prosody.convertToBI2Contour". We could toggle this feature with the following configuration setting:

```
prosody.convertToBI2Contour = true/false (default false)
```

**Note 2:** The examples given in this report will only work if the INPUT_TYPE is set to RAWMARYXML. HSMM voices are able to realize the specified prosody out of the box; for unit selection voices, at least MARY 4.3.0 is needed, and the checkbox "Apply prosody modification" must be activated in the web interface (http://mary.dfki.de:59125).
REFERENCES


