2.5. Methodology

Figure 2.2: Example of applying temporal smoothing to LSF parameterization using a sliding Hanning window. The width of the window was varied, to impose varying amounts of smoothing. Figure 2.2 shows an example of this process.

2.5.1.2 Variance scaling

Variance adjustment was implemented as a simple scaling of the standard deviation by a fixed factor. For each parameter (i.e., each LSF) in turn, the mean value over the utterance was found and subtracted before multiplying the parameter by a scalar value, and finally adding the mean back in. By altering the scalar value, the standard deviation is correspondingly adjusted, to simulate both reduced variance (which is commonly observed in HMM synthesis) and increased variance (e.g., as may happen if a Gaussian p.d.f. is poorly estimated during training, or when GV fails to re-instate the appropriate amount of variance). This approach of variance scaling is similar to the postfiltering method investigated by Silén and Helander (2012). Figure 2.3 shows an example of this process.
Unit selection

Independent Feature Formulation (IFF) target cost function
What you should already know

- selecting waveform fragments from a database of natural speech
- target cost
- join cost
- search
What you should already know

- selecting waveform fragments from a database of natural speech
- **target cost**
- join cost
- search

the target cost measures **mismatch** between

a target unit

and

a candidate unit
A target cost function based only on linguistic features

*The independent feature formulation (IFF)*

- Let’s start with the simplest form of target cost function

- It will simply **count** the number of **linguistic features** in the context of the candidate that **do not match** those of the corresponding target unit

- Motivation is simple
  - An exactly-matching candidate will have a cost of zero (= no mismatch)
  - The more mismatched the context is between candidate and target, the higher the cost

- The cost is a prediction of ‘how bad’ the candidate would sound, if used here
The IFF target cost function

Phonetic context

Stress

Syllable position

Word position

Phrase position
The IFF target cost function

Phonetic context
Stress
Syllable position
Word position
Phrase position

In the database, we have a recording of the sentence “A car.”
Festival's *multisyn* IFF target cost

<table>
<thead>
<tr>
<th>feature</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>stress</td>
<td>10</td>
</tr>
<tr>
<td>syllable position</td>
<td>5</td>
</tr>
<tr>
<td>word position</td>
<td>5</td>
</tr>
<tr>
<td>POS</td>
<td>6</td>
</tr>
<tr>
<td>phrase position</td>
<td>7</td>
</tr>
<tr>
<td>left phonetic context</td>
<td>4</td>
</tr>
<tr>
<td>right phonetic context</td>
<td>3</td>
</tr>
<tr>
<td><em>bad F0</em></td>
<td>25</td>
</tr>
<tr>
<td>duration outlier</td>
<td>10</td>
</tr>
</tbody>
</table>
Example calculation of IFF target cost for two competing candidates

<table>
<thead>
<tr>
<th>feature</th>
<th>weight</th>
<th>target</th>
<th>candidate 1</th>
<th>candidate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>stress</td>
<td>10</td>
<td>primary</td>
<td>primary</td>
<td>none</td>
</tr>
<tr>
<td>syllable position</td>
<td>5</td>
<td>coda</td>
<td>onset</td>
<td>coda</td>
</tr>
<tr>
<td>word position</td>
<td>5</td>
<td>final</td>
<td>final</td>
<td>final</td>
</tr>
<tr>
<td>POS</td>
<td>6</td>
<td>noun</td>
<td>noun</td>
<td>verb</td>
</tr>
<tr>
<td>phrase position</td>
<td>7</td>
<td>initial</td>
<td>initial</td>
<td>initial</td>
</tr>
<tr>
<td>left context</td>
<td>4</td>
<td>[b]</td>
<td>[b]</td>
<td>[v]</td>
</tr>
<tr>
<td>right context</td>
<td>3</td>
<td>[s]</td>
<td>[w]</td>
<td>[s]</td>
</tr>
</tbody>
</table>

**target cost =**
Another example, this time for **diphone** units

“Simon”

sil-s    s-ay    ay-m    m-ax    ax-n    n-sil
...time in...

...climbed...
Wait … how is prosody “created” using an IFF target cost function?

• With **no** explicit predictions of **any** acoustic properties, this is a reasonable question
• Answer:
  • candidates from appropriate contexts, when selected, will have appropriate prosody
  • the join cost will ensure that F0 is continuous

• So, we simply need to make sure the **linguistic features** capture sufficient contextual information that is relevant to prosody
  • e.g., stress status, position in phrase
• **Optional**: if our front end predicts **symbolic prosodic features** (e.g., ToBI accents and boundary tones), then we can use them in the target cost function
Unit selection

Acoustic Space Formulation (ASF) target cost function
Orientation

- Unit selection as we understand it so far
- run text processor (front end)
- construct target sequence
- retrieve candidates from database
  - compute IFF target costs
  - compute join costs
  - perform search

- Now, a more sophisticated target cost
  - predict **acoustic properties** of target units
  - compare these with actual acoustic properties of candidates
Orientation

- Unit selection as we understand it so far
  - run text processor (front end)
  - construct target sequence
  - retrieve candidates from database
    - compute IFF target costs
    - compute join costs
    - perform search

- Now, a more sophisticated target cost
  - predict acoustic properties of target units
  - compare these with actual acoustic properties of candidates
    - by comparing linguistic features
    - weakness: it is possible for two units with differing (mismatched) features to sound very similar
      - solution: compare how units sound
Figure 16.6

A diagram of four feature combinations lying in acoustic space, where only two dimensions of the high-dimensional acoustic space are shown for clarity. Note that, unlike in Figure 16.4, the positions of the feature combinations are not determined by the feature values, but rather by the acoustic definitions of each feature combination. Hence, these can lie at any arbitrary point in the space. In this case, we see that two feature combinations with quite different values lie close to each other, a situation that would not be possible in the IFF. The dotted ellipses indicate the variances of the feature combinations, which are used in some algorithms to measure distances.

16.4.1 Decision-tree clustering

The key part of the ASF is the design of a partial-synthesis function that can take any feature combination and map it onto the chosen acoustic space. The most common way of doing this is to use the decision-tree method, in more or less the same way as in HMM approaches (see Section 15.1.9). Since context accounts for a significant level of variance within a phone model, using separate phone models for each possible context greatly reduces the overall variance of the models. The problem faced, however, is that, while many of the required models have few or no observations in the training data, their parameters still have to be estimated. The similarity to our problem can now be seen: if some other acoustic property...
Predicting acoustic properties of the target units

• Think of this as ‘partial synthesis’

  • *do not* need to predict all acoustic properties
  • *do not* need to actually generate a speech waveform

  • just need to predict sufficient properties to allow comparison with candidate units
What exactly are the acoustic features?

- We have choices:
  - simple acoustic properties such as F0, duration and energy
  - a more detailed specification such as the spectral envelope (e.g., as cepstral coefficients)

- It will only work if we can **accurately predict** these properties from the linguistic features
  - how about predicting a **complete** acoustic specification?
Combining IFF and ASF into a single target cost function

• Many actual systems actually use a mixed IFF + ASF target cost function
  • some sub-costs use linguistic features, others use acoustic features
  • each is weighted appropriately

• Why use both types of sub-cost?
  • ASF escapes some of the sparsity problems inherent in IFF
  • but our acoustic properties do not capture all possible acoustic variation
    • e.g., voice quality, such as phrase-final creaky voice
  • and, of course, our predictions of acoustic properties will contain errors
Orientation

• Summary of unit selection design choices

• Unit type
• Target cost
• Join cost
• Search
• Database
Orientation

• Summary of unit selection design choices
  
  • Unit type
  • Target cost
  • Join cost
  • Search
  • Database
Orientation

- Summary of unit selection design choices

- Unit type
  Often diphones or half-phones.
  Use the "zero join cost trick" to effectively use (much) larger units

- Target cost
- Join cost
- Search
- Database
Orientation

• Summary of unit selection design choices

  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

  Pure IFF only using linguistic features
  Pure ASF, involving ‘partial synthesis’
  (must decide which acoustic features to predict)
  Mixed IFF + ASF
Orientation

- Summary of unit selection design choices
  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Usually includes F0, energy and spectral envelope

We have not mentioned optional smoothing of joins using signal processing.
Orientation

- Summary of unit selection design choices
- Unit type
- Target cost
- Join cost
- Search
- Database

Efficient dynamic programming

As in Automatic Speech Recognition, can use **pruning** to make it as fast as needed
Orientation

• Summary of unit selection design choices

• Unit type
• Target cost
• Join cost
• Search
• Database

Efficient dynamic programming

As in Automatic Speech Recognition, can use **pruning** to make it as fast as needed
Orientation

• Summary of unit selection design choices
  
  • Unit type
  • Target cost
  • Join cost
  • Search
  • Database

Coming next…
What next?

• How to create the database
• what to record
• how to record it
• how to annotate it

• Later, after we learn about statistical parametric speech synthesis
• we can use that statistical model in the ASF target cost function of a unit selection synthesiser
• this is called hybrid synthesis
What next?

• How to create the **database**
• what to record
• how to record it
• how to annotate it

Knowing what **features** our target cost requires, will help us design a suitable database of recorded speech.

• Later, *after* we learn about statistical parametric speech synthesis
• we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
• this is called **hybrid** synthesis
What next?

- How to create the **database**
- what to record
- how to record it
- how to annotate it

- Later, *after* we learn about statistical parametric speech synthesis
- we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
- this is called **hybrid** synthesis

We will have to annotate the database with the **features** that our target cost requires
What next?

• How to create the **database**
  • what to record
  • how to record it
  • how to annotate it

• Later, *after* we learn about statistical parametric speech synthesis
  • we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
  • this is called **hybrid** synthesis

Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan “A Unified Trajectory Tiling Approach to High Quality Speech Rendering” IEEE Trans. Audio, Speech, and Language Proc. 21 (2), pp. 280-290, 2013. DOI: 10.1109/TASL.2012.2221460