Modelling similarity perception of intonation

Uwe Reichel, Felicitas Kleber, Raphael Winkelmann
IPS, LMU München
reichelu|kleber|raphael@phonetik.uni-muenchen.de
25th June 2009
Content

- Introduction
- Perception of intonation similarity
- Relation between physical and perceptual intonation distance
- Modelling the perception of similarity
- Discussion and Conclusion
Introduction

Research context

- **Intonation modelling**
  - based on human perceptual equivalence judgements (e.g. *IPO*, t’Hart et al., 1990)
  - based on physical distance measures not motivated by human perception (e.g. *PaintE*, Möhler & Conkie, 1998)
  - **goal**: combine both $\rightarrow$ perceptual justification + automatisation

- **Evaluation of Speech synthesis systems** (Clark & Dusterhoff, 1999)

- **Second language acquisition** (Hermes, 1998)

Given Approaches

- **Physical measures**: e.g. correlation, absolute distance, RMS (Hermes, 1998), tangential and warping methods (Clark & Dusterhoff, 1999)

- **Evaluation**: e.g. correlation with human judgements derived from an ordinal scale (Hermes, 1998), so far up to 0.7.
Hypotheses and goals

- Ability of subjects to judge intonation similarity:
  1. Identical contours are judged to be more similar than different contours
  2. Contour judgements are consistent
- signal properties guiding the similarity judgements:
  3. There is a measurable relation between acoustic and perceived intonation similarity
Perception of intonation similarity

Subjects

- n=24 (17 female)
- age: from 20 to 42
- trained phoneticians: 19
- musical education: 14
- German native speakers: 19

Stimuli

- delexicalised [mama:ma] stimuli (vs. top-down processing)
- generated by Mbrola (male German voice; Dutoit et al., 1996)
- relevant f0 movement on the center syllable
- onset and nucleus durations: 60 and 200 ms, 130 and 300 ms, and 80 and 220 ms respectively (which was judged as natural and yielded the desired prominence relation in an informal pretest)
• **f0 generation:**
  – **target syllable:** third order polynomials, coefficients drawn randomly from ranges derived from f0 stylised corpus (IMS corpus, male German voice)
  – **remaining contour:** cubic spline extrapolation
  – **constraints:** concerning f0 range and distance of subsequent values

**Method**

• stimuli presented pairwise to the subjects over head phones (ISI: 0.5 sec, n(pairs)=300, presented once, 30 trial blocks)

• similarity judgement by clicking in a white area on the screen, the vertical position corresponding to perceived similarity

• no scale given since:
  – there is no sequence of equidistant categories related to similarity
  – ordinal scale hard to interpret (informal pretest)

• stimulus subsets:
  – IDENT: 20 pairs of identical contours to test **Hypothesis (1)**
  – CONSIST: 40 triplets (pairs presented 3 times) to test **Hypothesis (2)**

• removing judgement bias by normalising the answers to $[0 \ 1]$, reflecting the amount of perceived similarity
Results

- **Capability of similarity judgements**

![Box plots of similarity judgments](image)

Figure 1: **Left**: Perceived similarity of identical vs. differing contours. **Right**: Inconsistencies (standard deviations) for repeated pair and randomly combined pair triplets.

- means of identical vs. different contour similarity judgements: 0.92 vs. 0.43, h.s. (one-tailed Welch test, \( p < 0.0001 \)) \( \rightarrow \) **Hypothesis (1) confirmed**
- mean inconsistencies (standard deviations) of repeated vs. random pair triplets: 0.17 vs. 0.25, h.s. (one-tailed Mann-Whitney test, \( p < 0.0001 \)) \( \rightarrow \) **Hypothesis (2) confirmed**
- **Subjects are capable to judge intonation similarity**

Perception of intonation similarity
Relation between physical and perceptual intonation distance

- transforming similarity to distance judgements: \( d = 1 - s \)

**Correlations**

Table 1: *Pearson r* between perceived distance of intonation contours and a collection of their physical distances applied to raw f0 contours (in ST) and polynomial coefficients.

<table>
<thead>
<tr>
<th></th>
<th>contours</th>
<th>coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>0.40</td>
<td>0.38</td>
</tr>
<tr>
<td>Cityblock</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>Chebychev</td>
<td>0.47</td>
<td>0.38</td>
</tr>
<tr>
<td>1-Cosine</td>
<td>0.22</td>
<td>0.32</td>
</tr>
<tr>
<td>1-Correlation</td>
<td>0.33</td>
<td>0.29</td>
</tr>
</tbody>
</table>

- all correlations significantly different from zero (t-test, \( p = 0 \)) → **Hypothesis (3) confirmed**

- but nevertheless low → **metrics in isolation not capable of predicting perceived distance**
Relative weights

• grouping of the metrics by PCA loadings into four categories
  – $\text{pc}_1$: non-correlation-based distances for f0 contours
  – $\text{pc}_2$: non-correlation-based distances for polynomial coefficient vectors
  – $\text{pc}_3$: correlation-based distances ($1-\text{Cosine}, 1-\text{Correlation}$) of polynomial coefficient vectors
  – $\text{pc}_4$: correlation-based distances of f0 contours
• linear regression using $\text{pc}_1$–$\text{pc}_4$ as predictors for distance perception and comparing the regression weights
• result: $\text{pc}_1 > \text{pc}_3 > \text{pc}_2 > \text{pc}_4$
• **non-correlation-based distances of f0 contours have the highest relative influence on perceived distance**
Modelling the perception of similarity

Features

- 1−Correlation of the polynomial coefficient vectors
- pairwise absolute distances between the coefficient values
- Euclidean, Chebychev, and 1−Correlation distance between the onset contours (in ST) of the target syllable
- Euclidean, Chebychev, and 1−Correlation distance between the nuclei contours of the target syllable
- dichotomous algebraic sign comparison of the slope coefficients
- absolute differences in 7 equally sized area segments between the contours
- absolute difference of number of contour maxima
- previous answer of the listener

Preprocessing: orthogonalisation by PCA
Model 1: linear regression

- pairwise interaction model: $d_p = w_0 + \sum_i w_i f_i + \sum_i \sum_j w_{ij} f_i f_j$

Model 2: Two-layer feed-forward networks

Figure 2: Network Architecture
Figure 3: Neuron model. Output $O_i$ is given by its response $a$ to summed weighted input $I_i$. 

$$I_i = \sum_j w_{ij} O_j$$

$$O_i = a(I_i)$$
Figure 4: Activation functions $a(I_i)$. Here logsig is chosen.

- **training:**
  - modification of the weights $w_{ij}$ in order to yield outputs $d_p$ as close as possible to human distance judgements $d$
  - gradient descent backpropagation with momentum and adaptive learning rate vs. stranding in and oscillating around local optima
Figure 5: Gradient descent learning: update of weight $w_j$ guided by local minimisation of error $E$ (=MAE between $d$ and $d_p$).
Method

- excluding data from two subjects performing very badly with respect to judgement consistency
- 10-fold cross validation

Results

- **human performance:** standard deviation of the judgements for repeated contour pairs (= root mean squared error RMSE assuming, that the correct answer is given by mean value)

\[
\text{RMSE}_{d\text{-triplett}} = \sqrt{\frac{1}{3} \sum_{i=1}^{3} (d_i - \bar{d})^2}
\]

- **model performance:** RMSE for each model prediction (= absolute error)

\[
\text{RMSE}_{dp} = \sqrt{(d_p - d)^2} = |d_p - d|
\]
Figure 6: Human errors in terms of standard deviation of the judgement of repeated pair triplets. Absolute errors of the neural network and the regression model.

- one-way ANOVA, factor performer ("human" vs. "feed forward network" vs. "linear regression"): significant mean differences \( p = 0.002 \)
- Tukey-Kramer post-hoc: only significant differences between human and the linear regression performance
- trained feed forward networks do not perform significantly worse than the human listeners
Discussion and Conclusions

Setting of the perception experiment

- humans are able to perceive intonation similarities wrt judgement consistency
- worse performance of non-German natives: perhaps different prominence perception of center syllable
- not addressed yet:
  – longer segments than one target syllable
  – possible interference between perceptual similarity of two contours and their functional equivalence (Kohler, 1987)

Physical representation of perceived similarity

- low correlations for metrics in isolation and in combination
• **possible reasons:**
  – not all physical influence factors have been found yet
  – factors work together in a more sophisticated manner than examined here
  – the appropriateness of metrics is not adequately expressed in terms of correlation alone (see below)

• **further extensions:** e.g. weighting the contour distances by intensity (Clark&Dusterhoff, 1999)

• proposed method to determine the relative weight of influence factors by grouping them to PCs and by comparing the PC weights in a linear regression model

**Model evaluation**

• possible to develop acceptable feed forward network models to predict intonation distance

• performance not significantly worse than human performance vs. low correlation between model outputs and human perception data → suggesting that a model’s performance is not adequately expressed in terms of correlation alone
Acknowledgements

We would like to thank the participants of the seminar *Perceptive Phonetics* held in 2008/2009 at the University of Munich, who helped us to plan and to carry out the perception experiment.