INTERSPEECH 2009 BRIGHTON

Modelling similarity perception of intonation

Uwe D. Reichel, Felicitas Kleber, Raphael Winkelmann

Institute of Phonetics and Speech Processing
University of Munich
reichelu|kleber|raphael@phonetik.uni-muenchen.de

10.21437/Interspeech.2009-516

Abstract

In this study a perception experiment was carried out to examine the perceived similarity of intonation contours. Amongst other results we found, that the subjects are capable to produce consistent similarity judgements.

On the basis of this data we studied the influence of several physical distance measures on the human similarity judgements by grouping these measures to principal components and by comparing the weights of these components in a linear regression model predicting human perception. Non-correlation based distance measures for f0 contours received the highest relative weight.

Finally, we developed applicable linear regression and neural feed forward network models predicting similarity perception of intonation on the basis of physical contour distances. The performance of the neural networks, measured in terms of mean absolute error, did not differ significantly from the human performance derived from judgement consistency.

Index Terms: intonation, perception, similarity, neural networks

1. Introduction

The concept of intonation similarity is addressed in a variety of research fields ranging from intonation modelling [1] [2] over second language acquisition [3] to evaluation of speech synthesis systems [4].

Concerning the development of intonation systems, some approaches are based on human perceptual equivalence judgements, for example to manually adjust stylisations for original f0 contours in the IPO model [1]. Others utilise physical distance measures not motivated by human distance perception, e.g. for automatic intonation contour clustering as in [2] and [5].

It would be desirable to find objective measures based on human similarity perception which can be used for automatic similarity determination of intonation. Among the measures examined so far are correlation, absolute distance, and root mean squared distance between contours [3]. In [4] a tangential method is proposed to compute the contour distance orthogonally to the reference contour, as well as a warping method, in which only important contour segments that are to be specified in advance are compared.

The goodness of these measures is evaluated by correlation with human judgements usually derived from an ordinal scale [3] [4], or by calculating the ability of a distance measure d to separate cumulative relative frequency distributions of d values derived for each ordinal listener judgement level [3]. So far correlations up to a value of 0.7 have been reported.

1.1. Hypotheses and goals

The initial focus of our study was the question whether the subjects are at all able to judge intonation similarity. Regarding this, two hypotheses were formulated:

(1) Identical contours are judged to be more similar than different contours.

(2) Contour judgements are consistent.

If these hypotheses could be confirmed hypothesis (3) should reveal insight into the signal properities guiding the similarity judgements.

(3) There is a measurable relation between acoustic and perceived intonation similarity.

In section 2 of this paper a perception experiment to test hypotheses (1) and (2) is presented. In section 3 hypothesis (3) is tested and the relative influence of several acoustic distance measures on perceived similarity is examined. Finally, in section 4 we present applicable models predicting perceived similarity.

2. Perception of intonation similarity

2.1. Subjects

24 subjects (17 of them female) took part in the experiment. Their age ranged from 20 to 42. 19 subjects were trained phoneticians, and 14 subjects had a musical education. The mother tongue of 19 subjects was German. The non-German mother tongues each occurring once were: Italian, Spanish, Hungarian, Russian, and Slovenian. Two of the latter speakers have lived in Germany for over 10 years.

2.2. Stimuli

In order to reduce any kind of top-down processing, delexicalised [ma**ma:**ma] stimuli were generated by Mbrola (male German voice) [6]. The relevant f0 movement was placed on the centre syllable which was also lengthened in order to raise its prominence. The onset and coda durations for the three syllables were set to 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.

We generated the f0 contour of the target syllable by means of third order polynomials. The polynomial coefficient values were drawn randomly from a range derived from polynomial f0 stylisation of syllable segments in the IMS Radio News Corpus [7] (male German voice). The remaining contour in the carrier sequence was induced by cubic spline extrapolation. All contours had to fullfill the following constraints: their range had to be within the interval from 70 to 160 Hz, two subsequent values are not allowed to differ by more than 10%, and the f0 span within a syllable had to be less or equal to 50 Hz.

2.3. Method

The stimuli were presented to the subjects in pairs over head phones with a inter stimulus interval of 0.5 sec. The subjects' task was to judge the similarity of these pairs by clicking in a white area on the screen, the vertical position corresponding to perceived similarity. No scale was given to the subjects since we did not find any sequence of equidistant categories related to similarity, and in an informal pretest an ordinal scale turned out to be hard to interpret.

The experiment was comprised of 300 pairs regenerated for each subject as described in section 2.2 plus 10 initial stimulus pairs which served to get acquainted with the task (mean test duration: about 40 minutes). The stimuli were presented in randomised order without a repetition option. Within the presentation blocks of 30 trials, a subject's answer activated the next stimulus pair with a delay of 1 sec. After each block the subjects were able to decide when to continue.

To test hypothesis (1) a stimulus subset IDENT comprised of 20 pairs of identical contours was used. For hypothesis (2) another subset CONSIST was designed with 40 triplets each consisting of a contour pair presented three times in the course of the experiment.

To remove any judgement bias related to the used vertical span within the answering area, the answers were normalised to the interval $[0\ 1]$ reflecting the amount of perceived similarity.

2.4. Results

2.4.1. *Capability of similarity judgements*

In this study the subjects capability of perceiving similarity is reflected by hypotheses (1) and (2) concerning judgements of identical contours and judgement consistency.

Identical contours: Figure 1 shows the boxplots of the similarity judgements of identical contour pairs of the subset IDENT as opposed to differing contour pairs. The difference of the judgement means of 0.92 vs. 0.43 is highly significant (one-tailed Welch test, $p < 0.0001$). Hypothesis (1) can therefore be confirmed.

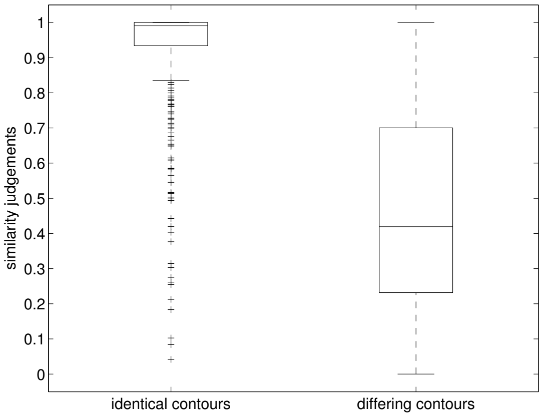

Figure 1: *Perceived similarity of identical vs. differing contours.*

Perception consistency: Judgement inconsistency can be expressed in terms of standard deviations. Figure 2 shows the overall difference in standard deviations between the repeated pair triplets of the subset CONSIST and an equally sized sample of randomly combined triplets. One can see, that the mean standard deviation for CONSIST triplets is lower than for random triplets (0.17 vs. 0.25). The difference is again highly significant (one-tailed Mann-Whitney test, $p < 0.0001$), i.e. the subjects were able to give relatively consistent judgements for repeated pairs. Hypothesis (2) can therefore be confirmed.

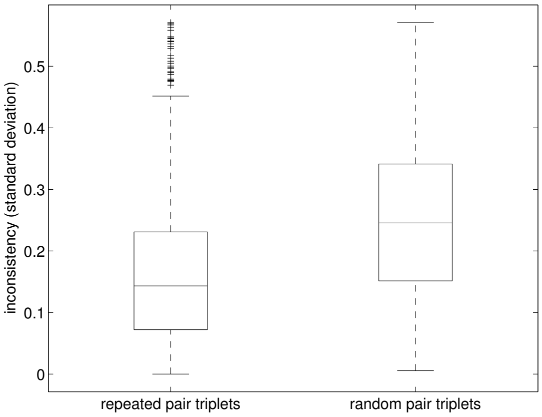

Figure 2: *Inconsistencies for repeated pair and randomly combined pair triplets.*

Performance differences of listener subgroups: Judgement consistency can be interpreted as a measure for judgement performance. The suchlike performance comparison of German and non-German mother tongue listeners yielded significant differences: Non-German natives performed significantly worse achieving a mean standard deviation of 0.22 as opposed to 0.17 (two-tailed Mann-Whitney test, $p = 0.002$). Nevertheless non-German natives still showed highly significantly higher consistency for the CONSIST subset reflected by standard deviations of 0.22 vs. 0.27 (two-tailed t-test, $p < 0.0001$).

Further performance differences were related to phonetic and musical training. Phoneticians and musicians performed significantly better than their respective counterparts (two-tailed Welch test, $p = 0.002$).

3. Relation between physical and perceptual intonation distance

For ease of comparison we transformed the subjects' similarity judgements s to distance judgements d as usual: $d = 1 - s$. Perceived distance was related to several distance measures to be found in table 1 together with their correlations to d. All distances were calculated for the polynomial coefficient vector pairs as well as for the generated f0 contour pairs on the target syllable. F0 was transformed from Hz to semitones, since the semitone scale is considered to be perceptually more relevant due to the results of a couple of studies (e.g. [1]).

All correlations are significantly different from zero (t-test, $p = 0$) supporting our hypothesis (3) stating a relation between acoustic and perceived intonation similarity. But this relation is nevertheless low (all $r < 0.5$). Therefore none of the proposed measures in isolation is capable of predicting the distance perception appropriately.

To examine the relative weights of the extracted distance measures, we grouped these measures by applying a principal

Table 1: *Pearson r between perceived distance of intonation contours and a collection of their physical distances applied to raw f0 contours (2nd column) and polynomial coefficients (3rd column).*

	contours	coefficients
Euclidean	0.40	0.38
Cityblock	0.38	0.37
Minkowski	0.40	0.38
Chebychev	0.47	0.38
1−Cosine	0.22	0.32
1−Correlation	0.33	0.29

component analysis explaining 98% of the variance. The result was a separation of the distance metrics into four groups (ordered as the associated principal components):

- pc_1: non-correlation-based distances for f0 contours
- pc_2: non-correlation-based distances for polynomial coefficient vectors
- pc_3: correlation-based distances (1−Cosine, 1−Correlation) of polynomial coefficient vectors
- pc_4: correlation-based distances of f0 contours

We then developed a linear regression model to predict the human distance judgements using the components associated with the 4 groups as predictors. The resulting absolute weights then reflect the influence of each of these feature groups on distance perception. pc_1 received the highest weight (0.0943) followed by pc_3 (0.0622), pc_2 (0.0475), and pc_4 (0.0053). Therefore in this study non-correlation-based distances of f0 contours had the highest relative influence on perceived distance. Nevertheless, all these weights are low in absolute terms as is the correlation between the linear model's output and the targets ($r = 0.47$).

4. Modelling the perception of similarity

4.1. Features

For each stimulus pair of the perception test the following 21 features were extracted from the data and related to the subjects' answers:

- 1−Correlation of the polynomial coefficient vectors
- pairwise absolute distances between the coefficient values
- Euclidean, Chebychev, and 1−Correlation distance between the onset contours 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

Some of the distance measures introduced in section 3 for whole contours and coefficient vectors are now applied on contour and coefficient vector segments to get a more detailed representation of the to be judged contour pair. None of these local features has a higher correlation than the global distance measures.

As above the contours are expressed in semitone values. Some subjects reported their impression, that some of their judgements were influenced by the preceeding answer. Altough in contrast the correlation between subsequent answers was low (r=0.14) we added this feature to the pool.

To remove the correlations among some of the features, they were orthogonalised by a principal component analysis. The subset of principal components explaining 99% of the total variance was then used as input for training the models.

4.2. Models

4.2.1. Linear regression

We utilised pairwise interaction models adding pairwise feature value multiplications to the linear additive terms.

4.2.2. Neural network

We trained and tested two-layer feed-forward networks consisting of one hidden layer with as many neurons as input features, and one output layer containing one neuron. All neurons were equipped with logarithmic sigmoid transfer functions. Training was carried out in 300 epochs by gradient descent backpropagation with momentum and adaptive learning rate against stranding in and oscillating around local optima respectively.

4.3. Method

In order to reduce the amount of noise in the training material, the data from two subjects performing very badly with respect to judgement consistency were excluded. Training and testing were carried out in form of 10-fold cross validation.

4.4. Results

4.4.1. Neural network vs. regression model performance

The performance of the models was measured in terms of correlation and of mean absolute errors between perceived and predicted intonation distance in the held-out data and is shown in Figure 3. The neural networks performed slightly better than the linear regression models yielding a mean correlation of 0.59 and a mean absolute error of 0.188 as opposed to 0.58 and 0.192 respectively, but this difference was not significant (two-tailed Mann-Whitney test, $p = 0.16$). It was not possible to increase the performance of the regression models utilising stepwise regression to remove irrelevant features.

4.4.2. Human vs. model performance

As described in section 2.4.1 the standard deviation of the judgements for repeated contour pairs is a possible measure for the human perceptual capabilities. Assuming that the correct answer of a judgement triplet is given by the triplet's mean value, standard deviation is equivalent to the root mean squared error of the human listener. To compare human and model performance we therefore employed the root mean squared error for each model prediction, which for single predictions corresponds to the absolute error, to compare it with the standard deviations derived from the subjects' answers.

Human mean error amounted 0.17, the models errors as already said around 0.19. A one-way ANOVA with the factor performer ("human" vs. "feed forward network" vs. "linear regression") revealed significant mean differences ($p = 0.002$).

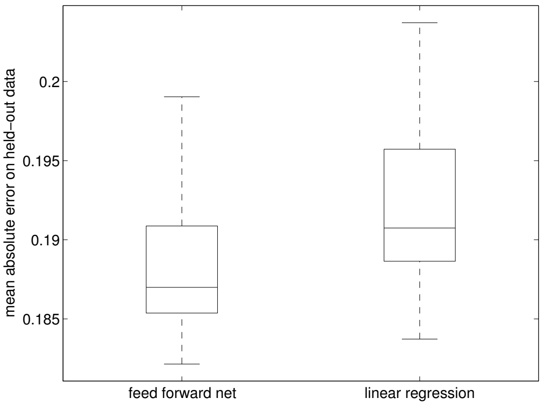

Figure 3: *Mean absolute errors of the neural networks and the regression models on the test data.*

According to the Tukey-Kramer post-hoc test the only significant difference could be found between the mean values of the human and the linear regression performance. This means that it cannot be concluded that the trained feed forward networks perform worse than the human listeners. The performances are shown in Figure 4.

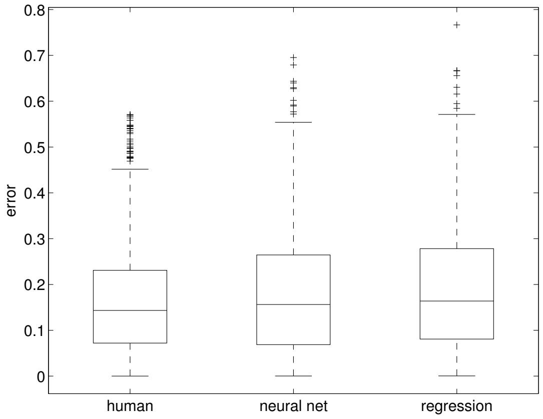

Figure 4: *Human errors in terms of standard deviation of the judgement of repeated pair triplets. Absolute errors of the neural network and the regression model.*

5. Discussion and Conclusions

5.1. Setting of the perception experiment

We were able to demonstrate that humans are able to perceive intonation similarities by using the concept of judgement consistency. Treating consistency as a performance criterion, a clear mother-tongue effect was observed, expressed in a worse performance of non-German natives. One of the reasons could be, that duration used to mark the target syllable is not in every language prominence-lending in such a degree as in German, so that non-German natives might have focused less on this syllable.

The setting had been restricted to just one target syllable. Future research has to reveal if the findings of this experiment can be generalised to longer segments.

Another important issue not addressed in this study is a possible interference between perceptual similarity of two contours and their functional equivalence. As has been shown in [8] continuous variation of intonation parameters can lead to non-linear shifts in the perceived function of the contour. It is not yet clear in detail how these shifts affect the perceptual distance of contours.

5.2. Physical representation of perceived similarity

All observed correlations between physical distance metrics and perceived distance turned out to be rather low. Also combining the measures by PCA and linear regression did not lead to increased correlations. This finding indicates that not all physical influence factors have been found yet and/or the factors work together in a more sophisticated manner than a simple linear combination. Further extensions of the feature pool could consist in e.g. weighting the contour distances by intensity as proposed in [4].

Nevertheless, in this study we proposed a method to determine the relative weight of influence factors by grouping them into principal components and looking at the weights of these components in a linear regression model of distance perception.

5.3. Model evaluation

It was possible to develop acceptable feed forward network models to predict intonation distance. Given a certain degree of variance within in the human judgements these networks did not perform significantly worse than humans. Since on the other hand the observed correlations between model outputs and the human perception data were not impressingly high, this finding further suggests, that a model's performance is not adequatly expressed in terms of correlation alone.

6. 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.

7. References

[1] J. t' Hart, R. Collier, and A. Cohen, *A perceptual study of intonation.* Cambrigde: Cambridge University Press, 1990.

[2] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proc. 3rd ESCA Workshop on Speech Synthesis*, 1998.

[3] D. Hermes, "Measuring the Perceptual Similarity of Pitch Contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, pp. 73–82, 1998.

[4] R. Clark and K. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proc. Eurospeech*, vol. 4, Budapest, 1999, pp. 1623–1626.

[5] U. Reichel, "Data-driven Extraction of Intonation Contour Classes," in *Proc. 6th ISCA Workshop on Speech Synthesis*, Bonn, 2007, pp. 240–245.

[6] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes," in *Proc. ICSLP*, vol. 3, Philadelphia, 1996, pp. 1393–1396.

[7] S. Rapp, "Automatisierte Erstellung von Korpora für die Prosodieforschung," Ph.D. dissertation, University of Stuttgart, Institute of Natural Language Processing, Stuttgart, 1998.

[8] K. Kohler, "Categorical pitch perception," in *Proc. ICPhS*, vol. 5, Tallinn, 1987, pp. 331–333.