Exemplar-based pitch accent categorisation using the Generalized Context Model

Michael Walsh, Katrin Schweitzer, Nadja Schauffler

Institute for Natural Language Processing, University of Stuttgart, Germany
<firstname>.<lastname>@ims.uni-stuttgart.de

Abstract

This paper presents the results of a pitch accent categorisation simulation which attempts to classify L*H and H*L accents using a psychologically motivated exemplar-theoretic model of categorisation. Pitch accents are represented in terms of six linguistically meaningful parameters describing their shape. No additional information is employed in the categorisation process. The results indicate that these accents can be successfully categorised, via exemplar-based comparison, using a limited number of purely tonal features.

Index Terms: pitch accent categorisation, exemplar theory, generalized context model, usage-based

1. Introduction

Exemplar-theoretic models of language processing have been shown to be suited for a variety of phenomena, such as vowel identification, sex identification [1], diachronic language change and frequency of occurrence effects [2], the emergence of grammatical knowledge [3, 4], syllable duration variability [5, 6], entrenchment and lenition [7], among others. While evidence for usage based production and perception is growing, as is the number of successful simulations of behavioural experiments by means of exemplar models, intonation has received little attention.

We present pitch accent categorisation simulation results, using a psychologically motivated exemplar model [8, 1], which demonstrate that exemplar-based categorisation of pitch accents is achievable with only a small number of tonal features to define each pitch accent.

The paper is structured as follows: section 2 describes the basic assumptions of exemplar theory. Section 3 argues for an exemplar-theoretic approach to intonation. Section 4 describes the Generalized Context Model, with which we carried out our simulations. Sections 5–7 detail our methodology and sections 8 and 9 provide an analysis of the results.

2. Exemplar theory

Exemplar theory [9, 1, 8, 7, 4] assumes that language is acquired in a usage-based fashion. That is, the main assumption is that concrete language input can be stored in memory in the form of "exemplars": single instances of previously perceived speech. Importantly, theses stored stretches of speech are assumed to be rich in detail – that is, unlike in abstract models of language, no normalisation processes, i.e. no loss of information, are assumed.

The accumulation of instances that are similar to each other is often termed "exemplar cloud", since it is assumed that similar instances are stored closely together in the exemplar space. Exemplars are assumed to be employed for both production and perception. In production, a production target is constructed from a set of exemplars. In perception, existing exemplars are used as references for categorisation. It is assumed that for newly incoming stimuli a similarity comparison takes place which determines the category of the stimulus, according to the properties of the previously encountered exemplars.

Computational models of the exemplar memory also argue that it is in a constant state of flux with new inputs updating it and old unused exemplars gradually fading away [7].

3. Exemplar theory and intonation

To date, to the authors' knowledge, no exemplar model exists which incorporates tonal parameters, nor do models that attempt to explain tonal phenomena work in usage based fashion. However, Hawkins and Smith [10] take a theoretical position which argues for a model that allows for storage of rich phonetic detail with pitch information included in the mental representation. Furthermore, there is evidence that tonal features should be part of the mental representation. Goldinger [9] reports a pilot study which found that subjects in a shadowing experiment adapt their pitch to the pitch of the speakers who recorded the stimuli – a finding that points to the possibility of pitch being an inherent part of the mental representation of a word. Results from production and perception studies clearly indicate that the mental representation of intonation must be rich in detail. For example, subjects in a mimicry task remember and reproduce fine detail of intonation [11]. Moreover, unfamiliar intonation contours slow down lexical access, indicating that intonation is directly involved when the word is accessed [12], although it is also suggested that the unusual contour might lead to interpretation problems and therefore a greater cognitive load which results in slower reaction times. Results from corpus analyses also point to a stronger coherence between words and their prosodies than traditionally assumed: Words often combine with the same tonal contour, and tonal contours can have different discourse meanings depending on the words they occur with [13]. In addition, combinations of words that occur together relatively often, display less variability with respect to their intonation than uncommon combinations of words [14].

Thus an increasing body of evidence seems to indicate that intonation contours can be stored, in rich detail, in memory. Consequently, they should be available to the speaker as production targets, and to the listener as exemplars against which new percepts can be compared. Given this, in the experiments that follow we employ Nosofsky's model of exemplar categorisation [8, 15] to simulate categorisation of intonation contours, or more specifically, pitch accents.

4. The Generalized Context Model

An exemplar-theoretic perspective on categorisation assumes that an incoming percept is categorised by comparing it against remembered exemplars of each potential category. The categorisation decision is made based on the extent to which the incoming percept activates each category, and this activation is modelled using sums of similarity over each category. The category to which the percept is considered most similar is the "winner". An important facet of the categorisation process is that not all features of the exemplars involved are considered equal, that is, certain features are attended to to a greater/lesser degree and play different roles in the decision making process.

In the context of the experiment described in section 7, the task is to categorise an incoming pitch accent with regard to two potential pitch accent types. Following [1] and [15] the model can be described and implemented using equations (1)-(4):

Given a percept i, and an extant exemplar j residing in memory, the similarity between them is calculated by first using equation (1) to determine the attention-weighted Euclidean distance d_{ij} between them, and then equation (2) to arrive at a similarity score using an exponential decay function. In equation (1), m represents an auditory property (in our case one of six features characterising the shape of the pitch accent) of i and j, and w_m is the attention weight given to feature m. In equation (2) c is a sensitivity constant which restricts distant exemplars from influencing the similarity score too much.

The attention-weighted Euclidean distance d_{ij} is given by:

$$d_{ij} = [\sum w_m (x_{im} - x_{jm})^2]^{1/2} \tag{1}$$

The auditory similarity s_{ij} is defined as:

$$s_{ij} = exp(-cd_{ij}) \tag{2}$$

Having arrived a similarity score for i and j, the extent to which percept i activates exemplar j in memory (assuming base activation N_j and optional Gaussian noise) is given by equation (3).

$$a_{ij} = N_j s_{ij} + e_j \tag{3}$$

The base activation reflects the fact that high frequency exemplar types are expected to have higher resting activation levels. The sum of activations of exemplars j of a particular category C_1 is then taken as evidence that percept i belongs to the category:

$$E_{1,i} = \sum a_{ij}, j \in C_1 \tag{4}$$

5. Categorising pitch accents

To see how the Generalized Context Model performs when used to categorise intonation parameters, we chose to categorise two pitch accent types. Pitch accents are tonal events that mark a syllable as prominent within the phrase. The data we used for our experiment is annotated manually with German ToBI pitch accent types according to the Stuttgart specification (GToBI(S), [16]). Since previous research has shown that inter-annotator agreement of pitch accent labelling is problematic [17] (in manual GToBI labelling the inter-transcriber consistency has been determined to be at only 51% for agreement on the full set of 9 different GToBI accent types) we chose two fairly distinct accents: a rising one, labelled as L*H in the GToBI(S) taxonomy and a falling one, H*L. According to the labelling scheme, these two accents are characterised by a low target in the accented syllable, followed by a rise on the post-accented one (L*H) and, respectively, by a high target in the accented syllable followed by a fall (H*L). In intonation labelling, however, annotators often rely more on their perceptual judgement than on pitch curves made visible by a fundamental frequency algorithm. For the database used in the study, annotators were specifically instructed to trust their perception rather the visualised contour, in cases of uncertainty.

As input for the exemplar-based categorisation model we used 6 parameters defining the shape of each accent. The parameters were retrieved using the PaIntE model which approximates stretches of smoothed F_0 contours. The approximation is carried out with the help of a mathematical function term with 6 free parameters. The function is built by summing up two sigmoids with a fixed time delay which is selected so that the peak does not fall below 96% of the function's range.

The sigmoids are subtracted from a basic value giving the function's maximum value within the analysis window. In this way the upper bound for the function is defined. The two sigmoids are defined each by 3 free parameters (a,b, and c, where a and c are sigmoid-specific and hence indexed according to their belonging to the first or the second sigmoid e.g. as $a1$ or $a2$, respectively) and a constant alignment parameter γ. The function term is given in equation (5).

$$\begin{aligned} f(x) = d - &\frac{c1}{1 + exp(-a1(b-x)+\gamma)} \\ - &\frac{c2}{1 + exp(-a2(x-b)+\gamma)} \end{aligned} \tag{5}$$

The 6 free parameters are linguistically motivated and reflect the shape of an accent (cf. figure 1): parameter b locates the peak of the accent within a three-syllable window, that is, the function normalises the syllable duration. Parameters *c1* and *c2* model the amplitudes of the rising and falling movement of the accent's contour, *d* corresponds to the absolute height of the peak and parameters *a1* and *a2* (not displayed in the figure) denote the "amplitude-normalised" steepness of the rising and falling slope (see [18, 19] for further information and illustrations concerning the mechanics of the PaIntE model).

Thus, the PaIntE parameters can be matched to the expectations for the shape of different GToBI(S)-accents in a straightforward way: H*L is defined as having a peak in the accented syllable, followed by a fall of the pitch contour. That is, the b parameter should be located in the accented syllable which means its value should be between 0 and 1 (in the temporally normalised syllable). As for the c-parameters, *c1* (the amplitude of the rise) would be expected to be small, whereas the amplitude of the fall, *c2*, would be expected to be high, since the definition of the accent requires a fall of the contour into the lower range of the speakers register. Likewise, for L*H, a higher b value would be expected (since there is a low target in the accented syllable, therefore a peak can only be in one of the post-accented syllables), and the values for *c1* would be expected to be high, whereas the *c2* values are expected to be small (cf. figure 2 which illustrates these assumptions by displaying the average shape of the two accent types in our data).

For our experiment, we used the newest version of the PaIntE model [20]. Note that the PaIntE model has different ways of approximating an accent: either both sigmoids of the function term are employed or, if the approximation is not successful in this way, only one of the accents is used (see [20] for details). This has implications for the data extraction (cf. section 6).

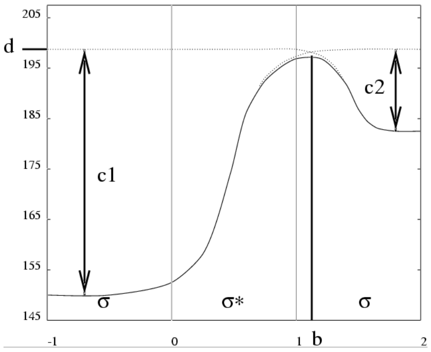

Figure 1: *The PaIntE model function is the sum of a rising and a falling sigmoid with a fixed time delay. The parameters are calculated over the span of the accented syllable ($\sigma*$) and its immediate neighbours. The x-axis indicates time (normalised for syllable duration) and the y-axis displays the fundamental frequency in Hertz.*

6. Data

For the categorisation task we used the DIRNDL corpus [21], which comprises recordings from three days of radio news broadcasts from a German radio station, summing up to more than 5 hours of read speech. The data was manually labelled and annotated for GToBI [16] pitch accent types. We represented each accent as a vector of 6 PaIntE parameters, describing the accent's shape.

We excluded those accents where the PaIntE model could only approximate the shape by deviating from the standard two-sigmoid case. This resulted in a dataset of 4604 H*L accents and 3787 L*H accents. The average accent shape for both accent types is displayed in figure 2.

Since the value distributions for the PaIntE parameters differ with respect to their means, ranges and standard deviations, we normalised each parameter by calculating the z-score (the number of standard deviations the respective value is away from the mean).

Importantly, we did not exclude outliers from the analysis in order not to make the categorisation task easier.

7. Experiments

One thousand tokens of each pitch accent type, L*H and H*L, were extracted from the corpus discussed above. The following experiments were then carried out:

7.1. Experiment 1 (a)

In a similar fashion to Johnson's vowel identification study [1], each L*H and H*L token was removed in turn from the corpus and treated as an incoming percept requiring categorisation using the L*H and H*L exemplar clouds and equations (1)-(4). As it is difficult to estimate the true frequency of occurrence of these pitch accent types, their base activation levels were both set to 1. Random noise was not employed. The attention weights (values between 0 and 1) for each of the PaintE parameters (w_{a1},w_{a2},w_b,w_{c1},w_{c2},w_d) were initially randomised and were then modified using simulated annealing to maximise categorisation accuracy. The sensitivity constant c was set to 0.105 following Johnson [1]. Given a L*H percept, if the evidence for membership of the L*H category exceeded the evidence for the H*L category, using equation 4, then the categorisation of the L*H percept was deemed correct. Categorisation of H*L tokens was performed in an analagous manner. The output is an accuracy rate over the input for a given set of attention weights. The annealing process iteratively modifies the weights and reruns the simulation in an attempt to increase the accuracy rate.

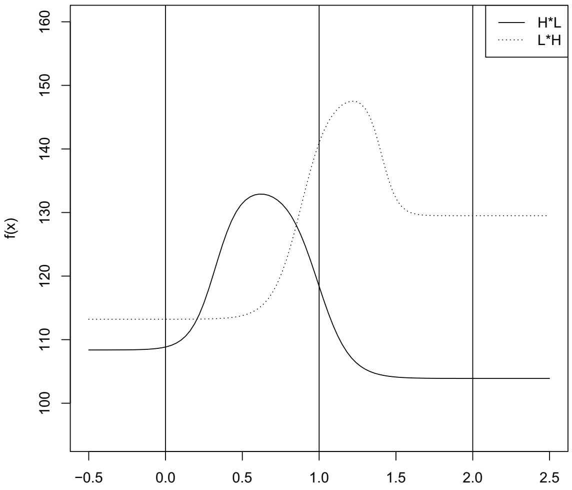

Figure 2: *Two accents (L*H and H*L) generated using the mean values for each PaIntE dimension in the employed subset of the DIRNDL database.The x-axis indicates time (normalised for syllable duration) and the y-axis displays the fundamental frequency in Hertz.*

7.2. Experiment 1 (b)

In order to test the robustness of the results from experiment 1 (a), a further 500 tokens of both L*H and H*L tokens were extracted from the corpus and categorisation was carried out, without annealing. The best-fitting weights from experiment 1 (a) were employed.

7.3. Experiment 2 (a)

To determine the contribution of those dimensions with the highest attention weights in categorising the accents, another categorisation experiment was carried out. In this experiment, we essentially repeated experiment 1 (a), using only the three dimensions that received the highest weights in experiment 1 (a), i.e. *b*, *c1* and *c2*, to categorise the two accents.

7.4. Experiment 2 (b)

To test the robustness of the results from Experiment 2 (a), we employed a similar methodology as in experiment 1 (b): a further 500 tokens of both L*H and H*L were extracted from the corpus and categorisation was carried out, without annealing, with the best-fitting weights from experiment 2 (a) employed.

7.5. Experiment 3 (a)

To investigate how well the model trained on PaIntE parameters performs compared to one trained on tonal features, which are directly derived from the F_0-contour, we run a further categorisation experiment in which we employed three tonal features

	Exp1	Exp2		Exp3
c	0.105	0.105		0.105
w_{a1}	0.40	–		
w_{a2}	0.40	–		
w_b	0.71	0.37	$w_{b'}$	0.66
w_{c1}	0.85	0.39	$w_{c1'}$	0.24
w_{c2}	0.99	0.46	$w_{c2'}$	0.69
w_d	0.15	–		
Accuracy (a)	80%	75%		68%
Accuracy (b)	81%	75%		66%

Table 1: *Attention weights and accuracy results from the experiments.*

which resemble the three PaIntE dimensions that had the highest weights in experiment 1.

The tonal features were derived using a speech synthesis system [22], which enabled us to encode the pitch of each syllable by means of 3 Hertz-values (extracted from the beginning, the middle and the end of the syllable). For each syllable we selected one pitch value out of these three as follows: If the pitch value for the middle of the syllable was not zero, we took this value (assuming that this should in most cases account for the pitch height within the nucleus). If, however, the algorithm did not find a valid value and thus returned zero, we took the mean value of the pitch at the beginning and at the end of the syllable. If one of those values was zero, as well, we took the remaining one. Thereby, we derived one pitch value for each syllable from which we then calculated, for each accented syllable, the difference in pitch height between the accented syllable and its preceding and following syllables. These two values roughly correspond to the PaIntE-values $c1$ and $c2$. To account for PaIntE-value *b*, i.e. the location of the peak, we determined the maximum pitch value of the three syllables and encoded its position as *1,2,* or *3* (corresponding to the pre-accented, the accented and the post-accented syllable). These three tonal features were z-scored and used as input to the Generalized Context Model to classify pitch accent types. These three parameters are termed $c1'$, $c2'$ and b' in table 1.

7.6. Experiment 3 (b)

This is analogous to experiments 1 (b) and 2 (b).

8. Results and discussion

Results from experiments 1 and 2 are presented in table 1. For experiment 1 (a), for the best-fitting set of weights, the ability of the model to categorise L*H and H*L accents is 80%. This is 30% above chance and multiple reruns of the simulation achieved similar results. Clearly, the model is reasonably capable of differentiating between the two accents, particularly in light of the fact that the low-level of inter-annotator agreement for pitch accent labelling would indicate that the data is probably somewhat noisy. Interestingly, the attention weights on which the model is most reliant are those for the PaIntE parameters which combine the most meaningful characteristics (though in different ways) of both pitch accent shapes: namely the location of the peak, $w_b = 0.71$, the amplitude of the rise, $w_{c1} = 0.85$, and the amplitude of the fall, $w_{c2} = 0.99$. The other parameters, which govern the steepness of the slopes and the overall height of the peak, while important in defining the shape of the accents, are intuitively not as important for discriminating between the two accents, and the exemplar model appears to pick up on this.

This is confirmed by our results from experiment 2 (a), which achieved 75% accuracy using only the three above dimensions. That is, the three highest-weighted PaIntE dimensions account for most of the classification, however, the additional features can improve the classification further. In addition, experiment 2 (a) confirms the ranking of the three parameters, with *c2* receiving the highest attention weight, followed by *c1*. Figure 2 demonstrates that, while the average L*H and the average H*L accent both have a relatively large amplitude of the rise (*c1*), the difference between *c2* for L*H and for H*L is greater, and the model identifies this by giving *c2* the greatest weight.

Experiment 1 (b), which tested the robustness of the results from experiment 1 (a), using a further 500 tokens of each pitch accent type and the weights reported in Table 1, produced slightly improved accuracy results: 81%. Analogously, experiment 2 (b), which tested the robustness of the results from experiment 2 (a), achieved an identical accuracy of 75%.

Experiment 3 (a), which analysed the categorisation performance of the model using three tonal features that were determined without using a sophisticated mathematical model function, but simply by comparing F_0 values, returned an accuracy of 68%. This demonstrates that while the classification is still considerably above chance, the more sophisticated PaIntE parameters contribute valuable information. The robustness testing (experiment 3 (b)) resulted in a slightly decreased accuracy of 66%, indicating that the model trained on these tonal features is consistently worse than the model trained on PaIntE features.

9. Conclusion

This paper lends further weight to the body of evidence advocating exemplar-based categorisation in language. To our knowledge this is the first attempt to successfully apply a psychological model of exemplar-based categorisation to pitch accents. Furthermore, the results show that an exemplar-based model of pitch accent categorisation can achieve reasonable and robust levels of accuracy on the basis of a very small number of linguistically motivated features which define the pitch accent shape.

It is important to note that the accents in the database were assigned by human annotators. Since they use far more information from the speech signal in the annotation task, e.g. spectral and temporal properties of the utterance, the model's achieved accuracy of 81% is reasonably high. Therefore one might speculate that the PaIntE parameters are valid approximations for dimensions in the perceptual space, as has been suggested elsewhere [20]. However, perception experiments would have to be carried out to further investigate this hypothesis.

Note that our experiments do not attempt to automatically annotate ToBI accents [23, 24, 25, 20], but rather seek to determine whether pitch accent categorisation can be captured by an exemplar-based psychological model of categorisation. Automated approaches to ToBI labelling tackle categorisation from an engineering perspective and use far more features.

Future work will also examine the effect of additional features, e.g. duration and segmental information, and explore categorisation using more accent types.

10. Acknowledgements

This research was supported by grants from the German Research Foundation (DFG): SFB-732, projects A1 and A2.

11. References

[1] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 145–165.

[2] J. Bybee, "From usage to grammar: The mind's response to repetition," *Language*, vol. 84, pp. 529–551, 2006.

[3] K. Abbot-Smith and M. Tomasello, "Exemplar-learning and schematization in a usage-based account of syntactic acquisition," *The Linguistic Review*, vol. 23, no. 3, pp. 275–290, 2006.

[4] M. Walsh, B. Möbius, T. Wade, and H. Schütze, "Multi-level exemplar theory," *Cognitive Science*, vol. 34, pp. 537–582, 2010.

[5] A. Schweitzer and B. Möbius, "Exemplar-based production of prosody: Evidence from segment and syllable durations," in *Speech Prosody 2004 (Nara, Japan)*, 2004, pp. 459–462.

[6] M. Walsh, H. Schütze, B. Möbius, and A. Schweitzer, "An exemplar-theoretic account of syllable frequency effects," in *Proceedings of the International Congress of Phonetic Sciences (Saarbrücken)*, 2007, pp. 481–484.

[7] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper, Eds. Amsterdam: Benjamins, 2001, pp. 137–157.

[8] R. M. Nosofsky, "Attention, similarity, and the identification-categorization relationship," *Journal of Experimental Psychology: General*, vol. 115, no. 1, pp. 39–57, 1986.

[9] S. D. Goldinger, "Words and voices—perception and production in an episodic lexicon," in *Talker variability in speech processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 33–66.

[10] S. Hawkins and R. Smith, "Polysp: a polysystemic, phonetically-rich approach to speech understanding," *Italian Journal of Linguistics - Rivista di Linguistica*, vol. 13, no. 1, pp. 99–188, 2001.

[11] B. Braun, G. Kochanski, E. Grabe, and B. S. Rosner, "Evidence for attractors in English intonation," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4006–4015, 2006. [Online]. Available: http://link.aip.org/link/?JAS/119/4006/1

[12] B. Braun, A. Dainora, and M. Ernestus, "An unfamiliar intonation contour slows down online speech comprehension," *Language and Cognitive Processes*, vol. 26, no. 3, pp. 350–375, 2011.

[13] S. Calhoun and A. Schweitzer, "Can intonation contours be lexicalised? Implications for discourse meanings," in *Prosody and Meaning (Trends in Linguistics)*, G. Elordieta Alcibar and P. Prieto, Eds. Mouton DeGruyter, 2012.

[14] K. Schweitzer, M. Walsh, S. Calhoun, and H. Schütze, "Prosodic variability in lexical sequences: Intonation entrenches too," in *Proceedings of ICPhS 2011*, Hong Kong, 2011, pp. 1778–1781.

[15] R. M. Nosofsky, "Exemplar-based accounts of relations between classification, recognition, and typicality," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 4, pp. 700–708, 1988.

[16] J. Mayer, "Transcribing German intonation – the Stuttgart system," Universität Stuttgart, Tech. Rep., 1995. [Online]. Available: http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html

[17] M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner, "Consistency in transcription and labelling of German intonation with GToBI," in *Proceedings of ICSLP*, 1996, pp. 1716–1719.

[18] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, pp. 311–316.

[19] G. Möhler, "Improvements of the PaIntE model for F_0 parametrization," Institute of Natural Language Processing, University of Stuttgart, Tech. Rep., 2001, draft version.

[20] A. Schweitzer, *Production and Perception of Prosodic Events – Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart, 2010.

[21] K. Eckart, A. Riester, and K. Schweitzer, "A discourse information radio news database for linguistic analysis," in *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, C. Chiarcos, S. Nordhoff, and S. Hellmann, Eds. Heidelberg: Springer, 2012, pp. 65–75.

[22] IMS Festival, "IMS German Festival home page," 2010, Insitut für Maschinelle Sprachverarbeitung, Universität Stuttgart. [Online]. Available: www.ims.uni-stuttgart.de/phonetik/synthesis

[23] A. Rosenberg, "AutoBI - a tool for automatic tobi annotation," in *Proceedings of Interspeech 2010*, 2010, pp. 146–149.

[24] V. K. R. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting Prominence in Conversational Speech: Pitch Accent, Givenness and Focus ," in *Proceedings of Speech Prosody (SP-2008)*, Campinas, Brazil, 2008, pp. 453–456.

[25] R. Fernandez and B. Ramabhadran, "Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Proceedings of Interspeech*, 2010, pp. 1429–1432.