Classification of Formal and Informal Dialogues Based on Turn-Taking and Intonation Using Deep Neural Networks

István Szekrényes[1] and György Kovács*[2,3]

[1] University of Debrecen, Debrecen, Hungary
[2] MTA Research Institute for Linguistics, Budapest, Hungary
[3] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
szekrenyes.istvan@arts.unideb.hu, gykovacs@inf.u-szeged.hu

Abstract. Here, we introduce a classification method for distinguishing between formal and informal dialogues using feature sets based on prosodic data. One such feature set is the raw fundamental frequency values paired with speaker information (i.e. turn-taking). The other feature set we examine is the prosodic labels extracted from the raw F0 values via the ProsoTool algorithm, which is also complemented by turn-taking. We evaluated the two feature sets by comparing the accuracy scores our classification method got, which uses them to classify dialogue-excerpts taken from the HuComTech corpus. With the ProsoTool features we achieved an average accuracy score of 85.2%, which meant a relative error rate reduction of 24% compared to the accuracy scores attained using F0 features. Regardless of the feature set applied, however, our method yields better accuracy scores than those got by human listeners, who only managed to distinguish between formal and informal dialogue to an accuracy level of 56.5%.

Keywords: Turn-Taking, Intonation, HuComTech, ProsoTool, Deep Neural Networks

1 Introduction

In the area of speech processing, spontaneous speech can be characterised in various ways. Many previous studies focused on the correlations of formal, measurable features and the underlying communicative or linguistic phenomena such as speech acts [15], topic structure [16, 21] and some paralinguistically relevant properties like age, gender and expressed emotions of the speakers [17]. Beyond the theoretical questions, the main practical challenge of these studies is how we can make the *content* – which is readily accessible for humans – machine-readable (detectable or predictable) based on physically measurable acoustic parameters.

The principle of our study is to characterise the situational context of dialogues by making a binary decision about the origin of topic units using neural

* The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant # K116938 and # K116402

nets and two kinds of dataset, namely formal and informal conversations taken from the HuComTech corpus [10]. Our assumption was that the different sequences of turn-taking (including overlapping speech and silence) and – based on the contextualizing function of prosody [7] – intonation movements described with normalised, discrete categories can provide sufficient information for a fairly successful classification of dialogue types. Following the approach of a previous study [11], the preprocessed annotation labels of turn-taking and intonation were used as training material instead of direct acoustic measurements, but for the sake of comparison, an experiment was also performed with the raw F0 data.

2 Research Material

The HuComTech multimodal corpus was designed within the framework of the HuComTech project [10] and used to analyse the underlying structure of human–human communication [9]. The corpus contains 50 hours of spontaneous speech in Hungarian recorded from 111 native speakers between the ages of 19 and 30. The speakers were asked to participate in a simulated job interview, and an informal conversation, discussing such topics as their happiest/saddest memories, friendship and jokes. Both scenarios were performed spontaneously and were directed by the same agent. Although it was only a simulation, the speakers were more polite and careful in their speech production during the job interview, than during the subsequent informal conversation. Because of the agent's directive role and the resulting unmotivated topic shifts, the two scenarios are very similar (making the classification more complicated), but based on the above-mentioned differences in behaviour and conversation topics, we were able to divide them into categorically different (formal and informal) subsets.

2.1 Annotation of Turn-Taking

In the HuComTech corpus, the speech of participants was transcribed manually in two separate annotation tiers, then the transcriptions were automatically converted to a simplified, acoustic representation of turn-taking that was divided

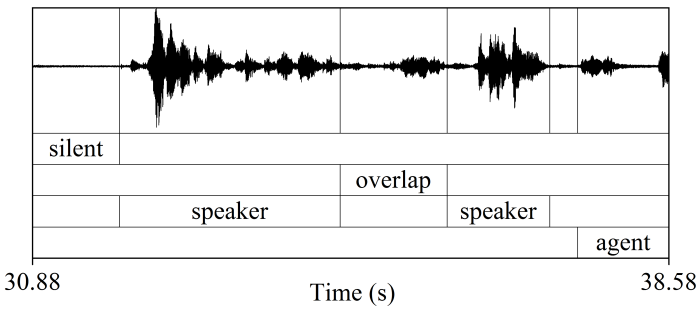

Fig. 1. Annotation of turn-taking. A sample taken from the HuComTech corpus.

Table 1. Average occurrences (per minute) and durations (in seconds) of speech segments (agent, speaker, overlapping speech) and silences

	Subset	Speaker	Agent	Overlap	Silence
Occurrence	Formal	9.25	5.59	2.48	10.43
	Informal	9.88	8.52	6.32	8.99
Duration	Formal	4.77 s	3.02 s	0.53 s	0.77 s
	Informal	3.09 s	2.55 s	0.76 s	0.67 s

into four levels: isolated speech segments of the *agent* and the *speaker*, segments of *overlapping speech*, and *silences*. As Figure 1 shows, the verbal interaction can be characterised by various patterns of consecutive events, using just these four categories of segmentation. In the absence of the original, manually created transcription, speaker diarisation algorithms are also available [8] to perform the task by means of automatic methods.

Information concerning the average occurrence (per minute) and average duration (in seconds) of various segments is displayed in Table 1. As can be seen, utterances of the agent are more frequent in the informal conversations than in the formal job interviews. This might be due to the fact that unlike in the job interviews, where the role of the agent was mostly limited to posing the initial questions and providing feedback to the interviewee (speaker), in the informal scenario the agent is more active and involved in the conversation. In Table 1, we can see a more significant difference regarding the frequency of overlapping speech. On average, overlapping speech is approximately 2.5 times more frequent in the informal conversations than it is in the formal job interviews. Although a higher frequency of overlapping speech would be expected with the increased activity of the agent, the increase in the frequency of overlapping speech is much bigger than the increase in the frequency of agent utterances. This means that the increased overlap in speech is probably a good indicator of the difference in the speaker behaviour in the two scenarios.

2.2 Annotation of Intonation

The annotation of intonation was performed by a rule-based algorithm called *ProsoTool*, which was implemented in the scripting language of the Praat speech processing program [1]. The development was inspired by the work of Piet Mertens [14] using similar objectives along with the psychoacoustic model of tonal perception. The main principle of ProsoTool was to transform the series of raw F0 values into smoothed, perceptually relevant, stylised trends of pitch modulation which can be classified as discrete contours of the evolving intonation structure. The script has a preprocessing module to isolate the voices of every speaker using the acoustic representation of turn-taking. Based on the F0 distribution, the algorithm divides the individual vocal range of speakers into five levels (see Figure 2), which were treated as normalised categories to locate the relative position of the resulting intonation trends. In Figure 3, the final

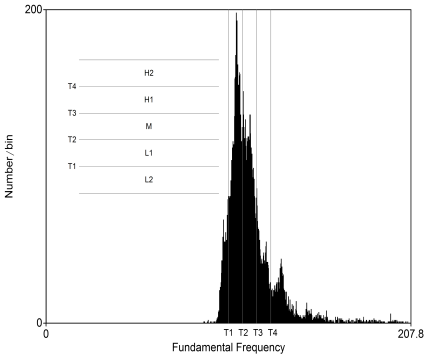

Fig. 2. Calculating individual pitch ranges of the speaker based on the F0 distribution: $L_2 < L_1 < M < H_1 < H_2$.

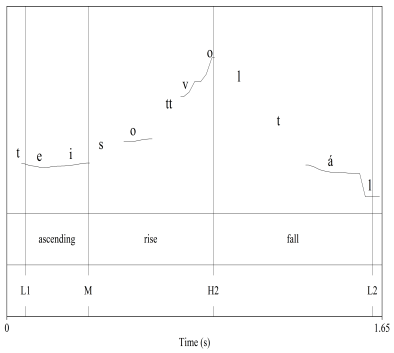

Fig. 3. Output for a Hungarian yes-no question: "Te is ott voltál? [Were you there, too?]".

output can be seen with the segmented and labeled F0 plots using five possible categories of intonation contour (*rise*, *fall*, *ascending*, *descending* and *level*), depending on the amplitude and the duration of the modulation. Further details on the method are described in [18]. The same algorithm is available at the official website[4] of the e-magyar project under the name of *emPros* (in accordance with the naming guideline of the project).

3 Methodology

3.1 Representation of data

Preparation and Conversion As a first preprocessing step, the annotations of conversations were divided into smaller units of dialogue topics. Some filtering criteria were also applied, namely pieces without any topic (e.g. the very beginning of conversations) or appreciable topic elaboration (when the total duration is less than 30 seconds) were excluded from the analysis. The annotation of turn-taking was converted to a sequential representation without any information on timing, but keeping the original order and marking the duration of each segment. In the case of ProsoTool's output, it was supplemented with the sequence of intonation trends describing the duration, the contour and the relative height of each trend (e.g. agent, 0.34, rise, M, H2). Unvoiced speech segments were also included to preserve the structure of turn-taking. In the third experiment, F0 values were measured everywhere using the default settings of the Praat "To Pitch..." function [1]. This resulted in sequences of 10-millisecond-long frames containing speaker information and the measured fundamental frequency in Hertz.

[4] `http://e-magyar.hu/en/speechmodules/empros`

Feature extraction The last step for converting our data into a format suitable for machine learning was feature extraction. Here, various different data types had to be handled, such as categorical data (speaker information and the F0 contour category), ordinal data (F0 level), and numerical data (the duration for each segment and raw F0 measurements). Categorical data was handled by 1 of N dummy coding (N being the number of categories), where silence was not considered as an independent category, but a lack of categorical information. It meant that turn-taking for example was coded as three binary features, corresponding to the three possible categories of speakers (speaker, agent, and overlapping speech), while all three binary features having the value of zero signified silence. The same method was used to encode the contour of the fundamental frequency (coding it as six binary features - corresponding to the five contour-types, and unvoiced intervals), as well as the ordinal data (also coded as six binary features). The only transformation applied on the numerical data was a standardisation to a zero mean and unit variance.

Train/Development/Test Partitioning To create separate sets to train our models, to tune the corresponding hyper-parameters, and also to evaluate the models trained, the speakers (and also the dialogue excerpts associated with the speakers) were partitioned into a train, development and test set. This partitioning was used in our auditory experiments as well, but this also meant that the potential size of the test set was limited by the workload we could realistically expect to be taken on by the volunteers in our experiments. In the end, similar to our earlier study on the HuComTech corpus [11], the partitioning was carried out with a 75/10/15 ratio. Thus from the 111 speakers, 17 were selected for the test set, 11 were selected for the development set, and the remaining 83 speakers were put into the train set. Both the test set and the development set were separated from the full set in such a way that they represented it as closely as possible. Meaning that from our candidate sets, we selected those whose parameters most closely resembled the full set. These parameters were the ratio of female/male speakers, the mean and deviation of the speakers' age, and the mean and deviation of formal and informal dialogue lengths. We also required that the number of formal and informal dialogues be equal in both the development set and the test set. This requirement helped to remove any unambiguity of the evaluation. And more importantly, the relative frequencies of informal and formal dialogues (the former slightly outnumbering the latter) in the full set of suitable dialogue-candidates of the HuComTech corpus did not necessarily reflect the real-life relative frequencies of such dialogues. As a consequence of our requirements, the partitioning resulted in a train set of 1058 dialogues, a development set of 136 dialogues, and a test set of 216 dialogues.

3.2 Machine Learning

Not only were there slightly more informal dialogues in the HuComTech corpus among dialogues suitable for our experiments, but they were also approximately

twice as long on average as their formal counterparts. We did not expect that this would be representative of formal and informal dialogues in general. Moreover, we sought to ensure that the classification should work regardless of the length of the dialogue, and regardless of whether the full dialogue was available for the classifier or not. For this, we decided to use a method that could not make use of the information of dialogue length. We applied a similar method to that used by Gosztolya [4] in the classification of laughter. For each segment with its context of a given size, a neural net estimated the probability of the given segment having been derived from an informal/formal dialogue, after which the classification was carried out based on the average of the resulting probability values.

Probabilistic sampling The difference between the average length of formal and informal dialogues also means that even though the number of formal and informal dialogues is roughly the same, at the segmental level there is a significant imbalance in the class distribution. This may cause a bias towards the more common (informal) class, and result in a worse classification performance of the rarer (formal) class [12]. One possible way of overcoming this problem is to omit entire informal dialogues, or just use smaller pieces of certain informal dialogues. This, however, would lead to the loss of important training data [2]. We might also try adding extra samples from the more rare class, or with the lack of extra examples, simulate this by using the same sample n times. In the probabilistic sampling method this can be achieved by first selecting a class at random, and then drawing a random sample from the selected class [20]. The first step can be viewed as sampling from a multinomial distribution, given that each class has a probability $P(c_i)$ [6]. That is,

$$P(c_i) = \lambda(1/N) + (1-\lambda)Prior(c_i) \qquad (1 \leq i, j \leq N; \lambda \in [0,1]), \qquad (1)$$

where N is the number of classes, and λ controls the uniformity of the distribution. Here, $\lambda = 0$ leads to the original distribution, while $\lambda = 1$ (a setting also referred to as "uniform class sampling" [20]) leads to a uniform distribution. In the second step, we take a random sample from the selected class.

Deep Rectifier Neural Nets Here, probability estimates for segments are provided by deep rectifier neural nets (DRNs). These are artificial neural nets that contain more than one hidden layer with neurons using the rectifier activation function ($rectifier(x) = max(0, x)$) instead of the traditional sigmoid function. As this architecture not only leads to more sparse neural nets, but also alleviates the problem of vanishing gradients even with multiple layers, it has gained popularity in recent years, not just in speech technology [5, 13, 19], but elsewhere as well [3, 6, 11]. The neural nets applied here had three hidden layers each containing 250 neurons, and an output layer containing two neurons, with a softmax nonlinearity. The training of the neural net was performed using the train set, while the development set was applied in the learn-rate scheduler, using Unweighted Average Recall (UAR) for validation.

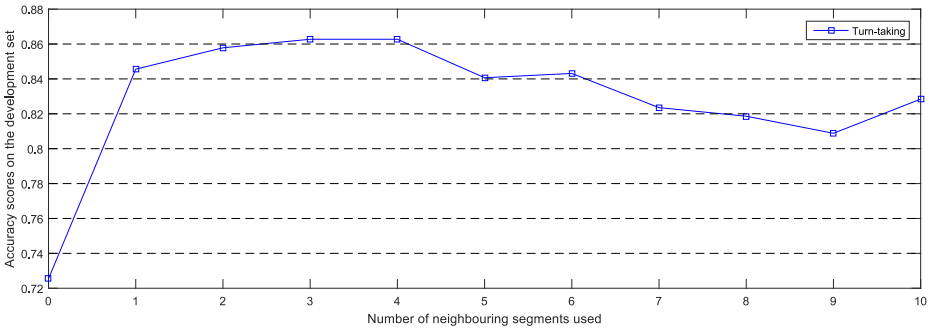

Fig. 4. Dialogue level accuracy scores got on the development set as a function of neighbouring segments used (the average of three independently trained classifiers)

4 Results and Discussion

4.1 Experiments using speaker information

In this study, one of our aims was to discover what classification accuracy could be attained using raw F0 measurements, and using information derived from these measurements with the ProsoTool algorithm. To make our feature sets more useful, both F0 measurements and ProsoTool labels were supplemented with turn-taking information (i.e. information on whether the current measurements correspond to an utterance of the speaker, the agent, both, or neither – if a measurement is taken during a period of silence). This raises the question of just how useful this information is in itself for classification purposes. We examined this question in our first set of experiments. Here, for each context size from 0 to 10 we trained three independent neural nets for $\lambda \in [0, 1]$. Figure 4 shows the average accuracy scores we got using the classifier with different neural nets. For each context size, the accuracy score of the best performing λ setting is shown. As can be seen in the figure, we achieved the best accuracy scores on the development set, when using 4-4 neighbours to estimate the probability values corresponding to a given segment. With this setting, using $\lambda = 0.9$, we obtained an accuracy score of 81.5% on the test set. This tells us that a reasonable classification performance can be achieved by just using turn-taking information as features.

4.2 Experiments using intonation

The same set of experiments was repeated using the feature set derived directly from the raw F0 data, as well as using the feature set derived from the output of ProsoTool algorithm. Figure 5 shows some results got from these experiments on the development set. As in the F0 feature set, one segment represents a much smaller context (10 ms); hence when using this filter set, more neighbouring segments were utilised in estimating the probability values of a segment derived

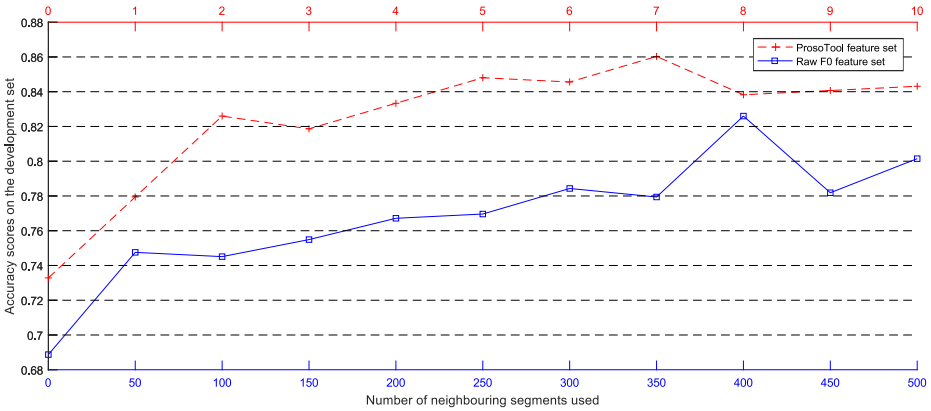

Fig. 5. Dialogue level accuracy scores got on the development set as a function of neighbouring segments used (the average of three independently trained classifiers)

from an informal/formal dialogue. Figure 5 shows that we get the best results using the raw F0 features, with a context of 400-400 neighbouring segments, while in the case of ProsoTool features, 7-7 neighbouring segments were used to obtain the best accuracy scores.

Table 2 shows the average accuracy scores got using the raw F0 and the ProsoTool feature sets, along with the average accuracy scores obtained using just the turn-taking information. As can be seen, utilising the F0 feature set (containing both the turn-taking information and the raw F0 measurements) not only failed to increase the accuracy scores compared to those obtained using just the turn-taking information, but even led to a slightly lower performance. This might seem counterintuitive, but the way the turn-taking information is represented within the F0 feature set (speaker information is given for every 10 milliseconds along with the fundamental frequency measurements), might be the reason it proved to be less efficient for the classifier. We also see that using the features from the ProsoTool algorithm did increase the performance of the classifier, leading to a relative error rate reduction of more than 24% compared to the accuracy scores got using the raw F0 feature set, and a relative error rate reduction of 20% compared to the classifier using just turn-taking information.

Table 2. Accuracy scores attained on the development set and test set using different feature sets (reported scores are the average of three independently trained classifiers)

Feature set	λ	No. of neighbouring segments used	Accuracy	
			Development	Test
Raw F0	1.0	400	82.6%	80.4%
Turn-taking	0.9	4	86.3%	81.5%
ProsoTool	1.0	7	86.0%	85.2%

To facilitate a comparison with the auditory experiments (see below), we created a classification based on the majority vote of the classifiers using the ProsoTool features. This resulted in an improved accuracy score of 85.6%.

4.3 Auditory experiments

Along with our machine learning experiments, an auditory experiment was devised to test the classification capability of human listeners based on the same information that was given to our automatic classification method, namely intonation and turn-taking. Hence, in this experiment, the original audio recordings (taken from the test set) were regenerated and presented as stereo channels (agent and speaker) of sine waves with varying frequency to represent only the intonation of communicative partners. The subjects had to listen to these samples through a Web-based interface and decide whether the dialogue was "formal" or "informal". Participants also had the chance to mark a dialogue as undecidable if they were unable to make a decision. The test set of 216 recordings was divided into three parts and three decisions were made about each recording by three different subjects. The final decision for each recording was made by a majority vote, resulting in an accuracy score of $56,5\%$. It should be mentioned here that the only training the human listeners received was an opportunity to familiarise themselves with the Web-based interface. Of course their performance could have been better if they had undergone a short training period where the correct answer was revealed after their decision made. However, as the aim of the experiment here was to determine their performance without any external assistance, this was not done.

5 Conclusions and future work

Here, we presented an algorithm for the classification of formal and informal dialogues based on intonation and turn-taking information. Despite the fact that the performance of human listeners on this task was generally not much better than what one would expect from a random decision, with our automatic classification method we achieved good accuracy scores. Furthermore, our results seem to confirm the utility of ProsoTool, as we achieved our best accuracy scores using the features provided by this algorithm.

In the future we would like to extend the dataset with the annotation of other prosodic features of speech rate and intensity using the upcoming, new modules of ProsoTool. And as the HuComTech is a multimodally annotated corpus, information from other modalities (such as facial expressions or deixis) could also be incorporated into the dataset. Besides this, it would be a good idea to examine more sophisticated methods for the aggregation of probability values. Here, this aggregation was carried out by a simple averaging, but it is not strictly necessary that every part of the dialogue should have the same importance regarding the final decision. We also intend to investigate different neural net architectures for this task, like Long-Short Term Memory (LSTM) neural networks and other recurrent networks.

References

1. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program]. version 6.0.22. http://www.praat.org/ (2016), retrieved 15/11/2016
2. Domingos, P.: A few useful things to know about machine learning. Commun. ACM 55(10), 78–87 (Oct 2012)
3. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proc. AISTATS. pp. 315–323 (2011)
4. Gosztolya, G., Beke, A., Neuberger, T., Tóth, L.: Laughter classification using Deep Rectifier Neural Networks with a minimal feature subset. Archives of Acoustics 41(4), 669–682 (2016)
5. Grósz, T., Busa-Fekete, R., Gosztolya, G., Tóth, L.: Assessing the degree of nativeness and Parkinson's condition using Gaussian Processes and Deep Rectifier Neural Networks. In: Proc. Interspeech. pp. 1339–1343 (2015)
6. Grósz, T., Nagy, I.: Document classification with deep rectifier neural networks and probabilistic sampling. In: Proc. TSD. pp. 108–115 (2014)
7. House, J.: Prosody and context selection: A procedural approach. In: Barth-Weingarten, D., Dehé, N., Wichmann, A. (eds.) Where Prosody Meets Pragmatics, pp. 129–142. Emerald (2009)
8. Huijbregts, M.: Segmentation, diarization and speech transcription: surprise data unraveled. Ph.D. thesis, University of Twente (2008)
9. Hunyadi, L.: Multimodal human-computer interaction technologies. Theoretical modeling and application in speech processing. Argumentum pp. 240–260 (2011)
10. Hunyadi, L., Váradi, T., Szekrényes, I.: Language technology tools and resources for the analysis of multimodal communication. In: Proc. LT4DH, pp. 117–124. University of Tübingen, Tübingen (2016)
11. Kovács, Gy., Grósz, T., Váradi, T.: Topical unit classification using deep neural nets and probabilistic sampling. In: Proc. CogInfoCom. pp. 199–204 (2016)
12. Lawrence, S., Burns, I., Back, A., Tsoi, A.C., Giles, C.L.: Neural network classification and prior class probabilities. In: Orr, G.B., Müller, K.R. (eds.) Neural Networks: Tricks of the Trade, pp. 299–313. Springer, Berlin, Heidelberg (1998)
13. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML. vol. 30 (2013)
14. Mertens, P.: The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In: Proceedings of Speech Prosody (2004)
15. Mushin, I., Stirling, L., Fletcher, J., Wales, R.: Identifying prosodic indicators of dialogue structure: Some methodological and theoretical considerations. In: Proc. SIGdial. pp. 36–45. Association for Computational Linguistics (2000)
16. Nakajima, S., Allen, J.F.: A study on prosody and discourse structure in cooperative dialogues. Tech. rep., Rochester, NY, USA (1993)
17. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in speech and language-state-of-the-art and the challenge. Comput. Speech Lang. 27(1), 4–39 (Jan 2013)
18. Szekrényes, I.: ProsoTool, a method for automatic annotation of fundamental frequency. In: Proc. CogInfoCom. pp. 291–296 (2015)
19. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proc. ICASSP. pp. 6985–6989 (2013)
20. Tóth, L., Kocsor, A.: Training HMM/ANN hybrid speech recognizers by probabilistic sampling. In: Proc. ICANN. pp. 597–603 (2005)
21. Zellers, M.: Prosodic variation for topic shift and other functions in local contrasts in conversation. Phonetica 69(4), 231–253 (2013)