Automatic Detection of Hyperarticulated Speech

Eugénio Ribeiro[1,2], Fernando Batista[1,3], Isabel Trancoso[1,2], Ricardo Ribeiro[1,3], and David Martins de Matos[1,2]

[1] L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] ISCTE-IUL – Instituto Universitário de Lisboa, Portugal
eugenio.ribeiro@l2f.inesc-id.pt

Abstract. Hyperarticulation is a speech adaptation that consists of adopting a clearer form of speech in an attempt to improve recognition levels. However, it has the opposite effect when talking to ASR systems, as they are not trained with such kind of speech. We present approaches for automatic detection of hyperarticulation, which can be used to improve the performance of spoken dialog systems. We performed experiments on Let's Go data, using multiple feature sets and two classification approaches. Many relevant features are speaker dependent. Thus, we used the first turn in each dialog as the reference for the speaker, since it is typically not hyperarticulated. Our best results were above 80% accuracy, which represents an improvement of at least 11.6 percentage points over previously obtained results on similar data. We also assessed the classifiers' performance in scenarios where hyperarticulation is rare, achieving around 98% accuracy using different confidence thresholds.

Keywords: hyperarticulation · speech · Let's Go

1 Introduction

When people face recognition problems by their conversational partners, they tend to adopt a clearer form of speech in an attempt to improve recognition levels. This speech adaptation is called hyperarticulation. However, although it may work in child-directed speech or when talking to people with hearing impairment, it typically has the opposite result when talking to an Automatic Speech Recognition (ASR) system, decreasing its performance [17, 7]. This happens because ASR systems are not trained with hyperarticulated speech. Furthermore, these situations typically occur in attempts to recover from previous recognition errors. This means that the supposed correction will also be misrecognized, leading to further hyperarticulation and recognition errors, completely disrupting the dialog flow [11].

Automatic detection of hyperarticulated speech is important because if a dialog system is able to do so, it can try to recognize the utterance using ASR models trained with hyperarticulated speech, or at least try to guide the user towards the use of unmarked speech. Furthermore, automatic hyperarticulation

detection can be used to find possible causes of hot spots in a dialog, reducing the need for manual annotations.

Our experiments on automatic detection of hyperarticulation were performed on data from real user interactions with an Interactive Voice Response (IVR) system along multiple years. In this paper we present several approaches to the task, using multiple feature sets and classification algorithms.

The remaining sections are structured as follows: Section 2 presents the related work. Section 3 describes the datasets, features and approaches used. Results are presented and discussed in Section 4, and, finally, Section 5 states the conclusions and suggests paths for future work.

2 Related Work

Hyperarticulation is widely accepted as an important factor that should be dealt with by speech applications. However, not much effort has been put into its automatic detection. Research involving hyperarticulation is usually directed towards the evaluation of its impact in ASR and how systems can adapt to it or redirect the users towards the use of unmarked speech. Nonetheless, there are multiple studies that explore the characteristics of hyperarticulated speech. This is important for the automatic detection of hyperarticulation as those characteristics help selecting relevant features for the task.

The work by Fandrianto and Eskenazi [5] on prosodic entrainment of shouting and hyperarticulation is, to our knowledge, the only one that provides concrete results for automatic hyperarticulation detection. The automatic detection approach consists of a Support Vector Machine (SVM) classifier trained using Let's Go [14] data from the years of 2009 and 2010 that had been used in previous experiments [12]. In terms of features, the authors used a small set of six acoustic features – fundamental frequency range and average, intensity, harmonic-noise ratio, and pause duration and frequency –, extracted using openSMILE [4], and two dialog-level features – ASR confidence and explicit confirm repetition. Using this approach, the authors achieved 70% accuracy on a balanced test set.

Oviatt et al. [11] studied hyperarticulated speech and analyzed how a set of features changed when people started to hyperarticulate. The analyzed features involved durations of both speech and pauses, speech rate, amplitude, fundamental frequency, intonation contour, phonological alternations, and disfluencies. The authors reported a significant increase in both the number and duration of pauses during hyperarticulated speech. On the contrary, the number of disfluencies reduced significantly. Also, the final fall in intonation, speech elongation, and the use of hyper-clear phonology moderately increased. Finally, minimum and average pitch slightly decreased.

Soltau and Waibel [18] analyzed hyperarticulated speech at the phone level in order to develop acoustic models able to deal with that kind of speech. By analyzing vowel formants, the authors only found significant differences for the vowel /uw/. In terms of duration, all phones lasted longer during hyperarticulated speech. However, the difference for consonants was two times the one

for vowels. Also, the duration of voiced plosives increased over 40%. Finally, in terms of the place of articulation, retroflex and labiodental sounds revealed no dependence of hyperarticulated speech, while velar, alveolar, and bilabial sounds revealed at least some level of dependence.

Stent et al. [19] also studied hyperarticulation at the phone level. The authors paid special attention to the /t/ phone, specially when it comes before an unstressed vowel, at the end of the word, or after an *n*, as its pronunciation differs during clear speech. The presence of a full vowel in the definite article *a* and of the /d/ phone in *and* was also analyzed. However, in these cases, the studies showed no clear differences, which, according to the authors, may be explained by the tendency to use clearer speech on content words rather than on function words when repairing, as content words are more critical to understand a message. Furthermore, the authors analyzed the vowel formants and concluded that front vowels become even more fronted during hyperarticulated speech.

Overall, we can conclude that hyperarticulated speech is characterized by changes in multiple acoustic-prosodic features, as well as articulatory changes in specific phones.

3 Experimental Setup

This section describes our experimental setup, starting with the adopted datasets, followed by the used feature sets, classification approaches, and evaluation methodology.

3.1 Datasets

In our experiments we used data from three years of The Let's Go corpus. The corpus features data from the CMU Let's Go Bus Information System, which provides information about bus schedules in the city of Pittsburg, through spoken telephonic interaction with a dialog system. Subsets of the data from the three years were annotated for hyperarticulation in a joint effort by L^2F and KTH [8]. 834 turns from 2009, 1110 from 2012, and 1449 from 2014 were annotated, out of which 113, 90, and 77, respectively, were labeled as hyperarticulated. It is important to notice that the system evolved over the years, both in terms of ASR performance and dialog management. Thus, the characteristics of the data change according to the year. Since the datasets are highly unbalanced, we balanced them using the Spread Subsample filter provided by the Weka Toolkit [6] to obtain datasets with the same number of examples of each class. We performed experiments using each of the balanced yearly datasets individually, as well as together in an aggregated dataset. Furthermore, we used the unbalanced datasets for precision evaluation.

3.2 Features

Taking the characteristics of hyperarticulated speech mentioned in Section 2 into account, we extracted sets of acoustic, segmental, and disfluency-based features

that intend to capture some of those characteristics. Furthermore, we used the ComParE 2013 feature set, since it is widely used in speech-related tasks.

ComParE 2013 The ComParE 2013 feature set [16], extracted using openSMILE, provides a large amount of acoustic-prosodic features and is widely used in speech analytics tasks. Hyperarticulation is inherently related to acoustic and prosodic factors, thus, this set may be suitable for this task.

Segmental Hyperarticulation is highly related to rhythmic and durational features. The extraction of such features requires segmentation of the original audio file into smaller, informative segments. We obtained a phone tokenization of the audio using the neural networks that are part of our in-house ASR system [9].

From the segmentation directly, we extracted features related to phones – count, total speech duration, average phone duration, speech ratio, and phone-based speech rate – and pauses – count, total silence duration, average pause duration, silence ratio, and silence-to-speech ratio. In terms of Inter-Pausal Units (IPUs), i.e., sequences of phones between two pauses, we extracted the count, rate, and 9 statistics – maximum, minimum, mean, standard deviation, median, mode, slope, concavity, and range – of their duration, in seconds, and length, in number of phones. The IPUs were also important for the extraction of acoustic features, as described below.

Acoustic In terms of base acoustic features, we extracted energy, pitch, and Harmonic-Noise Ratio (HNR) using openSMILE, with overlapping windows of 50ms, and a 10ms step. We also extracted normalized amplitude using SoX[4]. We computed the same 9 statistics listed in the previous section for each of the features, using the whole audio file. Furthermore, we computed the same statistics for the data corresponding to each IPU in the audio file, and repeated the same procedure to obtain IPU-based statistics for the whole file. Finally, we also computed the same 9 statistics for pitch, discarding null values, that is, those corresponding to unvoiced speech or non-speech.

Disfluencies As stated in related work [11], the number of disfluencies tends to decrease during hyperarticulated speech. Thus, we counted the number and calculated the ratio of IPUs that contained disfluencies. We took advantage of the speech disfluency detection module [10] provided by the SPA[5] [1] speech analytics platform. However, it is important to notice that this module was trained using data from the CORAL [20] corpus, which contains non-English data.

[4] `http://sox.sourceforge.net`
[5] `https://www.l2f.inesc-id.pt/spa/`

First Turn Differences Most of the characteristics of hyperarticulated speech are speaker dependent. For instance, when hyperarticulated speech is characterized as slower, it is slower in relation to the normal speech rate of the person. Since our dataset consists of turns extracted from human-machine dialogs, we are able to extract all the previously described features from the first human turn of each dialog. That turn is highly unlikely to contain hyperarticulated speech, as the person has not faced recognition problems by the machine. Thus, it can be used as a reference for the remaining turns in the dialog. Taking advantage of this, we computed the difference between the value for each of the previously described features and the corresponding value for the first turn in the same dialog.

3.3 Classification

In our experiments, we considered the detection of hyperarticulated speech a binary classification task. From the multiple classification approaches that could be used, we opted for SVMs [3], which are widely used and typically produce acceptable results, and Random Forests (RFs) [2], an approach based on decision trees, which has been proved successful in experiments using similar data [15].

To train our SVMs, we took advantage of the Sequential Minimal Optimization (SMO) algorithm [13] implementation provided by the Weka Toolkit. We used the linear kernel and kept the C parameter with its default 1.0 value, as it led to the best results in our experiments. We also took advantage of the Weka Toolkit to train our RFs. Since the number of instances is small, it is affordable to generate a large amount of trees. Thus, we used 1000 as the value of that parameter.

3.4 Evaluation

In order to evaluate our approaches, we used two different procedures. One that evaluated the importance of different feature sets for the task and the performance of the different classification approaches, and one that evaluated the capabilities of the approaches to adapt to real situations, with unbalanced datasets.

For the first, we used 10-fold cross-validation on both each yearly balanced dataset, as well as on the aggregated dataset. In this case, given the use of balanced datasets, the binary nature of the task, and the objective of evaluating the overall performance of the different feature sets and classification approaches, we relied solely on accuracy as evaluation measure.

For the second procedure, we used all data from each year, in order to simulate a real situation, with rare occurrences of hyperarticulated speech. In this case, in addition to accuracy, we also looked into precision, recall, and F-measure, to identify the best confidence threshold that reduced the number of false positives without highly increasing the number of false negatives.

4 Results

In this section we present the results obtained by the SVM and RF approaches using the different feature set combinations.

4.1 Balanced Datasets

Table 1 presents the accuracy results obtained by the SVM and RF approaches on the four datasets, using different feature sets. The table is split into three subtables, which contain the results obtained using the ComParE 2013 feature set, our acoustic-prosodic features, and the combination of the two. The rows labeled as **All** refer to the results obtained using all features of each subtable's category. The sets labeled as **Selected** were obtained by applying the Best First feature selection algorithm, with five consecutive nodes without improvement as the stop criterion, to the corresponding **All** set.

Table 1. Accuracy results obtained using the SVM and Random Forest approaches and the different feature sets.

	Let's Go 2009		**Let's Go 2012**		**Let's Go 2014**		**All**	
Feature Set	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
ComPaRe 2013								
Current Turn	0.668	0.717	0.761	0.750	0.727	0.773	0.689	0.748
1st Turn Difference	0.721	0.730	0.761	0.778	0.721	0.760	0.704	0.723
All	0.695	0.730	0.800	0.794	0.792	0.825	0.723	0.771
Selected	**0.805**	**0.841**	**0.839**	**0.872**	**0.818**	**0.935**	**0.766**	**0.805**
Acoustic-Prosodic								
Amplitude	0.664	0.695	0.572	0.678	0.623	0.721	0.613	0.723
Energy	0.708	0.712	0.633	0.683	0.662	0.747	0.673	0.713
Pitch	0.655	0.655	0.656	0.661	0.649	0.675	0.639	0.679
HNR	0.606	0.686	0.611	0.672	0.682	0.682	0.613	0.675
Pitch + Energy	**0.730**	0.730	**0.683**	0.672	0.649	0.766	0.671	0.723
Acoustic	0.690	0.735	0.622	0.711	0.662	**0.812**	0.684	0.730
Segmental	0.664	0.646	0.661	0.667	0.675	0.721	0.666	0.673
Disfluencies	0.566	0.571	0.600	0.617	0.468	0.552	0.555	0.559
P + E + H + Sil [5]	0.655	0.677	0.611	0.711	0.636	0.695	0.655	0.718
All	0.637	0.708	0.622	0.689	0.591	0.721	0.670	0.725
Selected	0.695	**0.739**	0.622	**0.712**	**0.766**	**0.812**	**0.718**	**0.738**
Combination								
All	0.699	0.735	0.772	0.783	0.773	0.792	0.720	0.764
Selected	**0.827**	**0.836**	**0.844**	**0.861**	**0.812**	**0.922**	**0.761**	**0.816**

The first important point to notice is that the RF approach systematically obtained better results than the SVM approach. This suggests that at least some of the features follow a distribution that is highly discriminative for the hierarchical structure of decision trees. Thus, the following remarks will be based on RF results.

Starting with ComPare 2013 features, we can see that this feature set on its own obtained accuracy results over 70% on every dataset, which defines a relatively high baseline, already above the one defined by Fandrianto and Eskenazi [5]. The use of feature value differences between the current turn and the first turn in the dialog improved the baseline results on every dataset except the one from 2014. Furthermore, the combination of the two sets, **All**, improved the results on every dataset. This proves the importance of the relation between the current turn and the first in the dialog, reducing the effects of speaker dependence. Performing feature selection on such a large set of features is very important, as many features provide no information. This can be proved by the results obtained by the **Selected** feature set, which are above 80% on every dataset.

In terms of our acoustic-prosodic features, the results presented for each set already contain the features related to the value differences in relation to the first turn, since those features were always able to improve the results. Starting with acoustic features, energy features were typically the most informative. The combination of pitch and energy, a typical combination when acoustic features are used, was able to achieve at least the same result as the best individual acoustic feature class. Furthermore, by appending the remaining two classes, the results improved for every dataset, surpassing the baseline for the 2009 and 2014 datasets. Segmental features were less informative, with accuracy results below each single acoustic class. However, this can be explained by the disappointing results obtained by some of the features in that set. For instance, speech rate only achieved results below 60%. However, there were also positive results, obtained by features based on durations, length, silence and speech. Still, in this sense, durational features were more informative than the ones related to length and the same happened with speech-based features in relation to silence-based features. Disfluency-based features also obtained disappointing results, below 60% accuracy. However, in this case, the results can be explained by the fact that the disfluency detector was trained using data in a different language. The combination of pitch, energy, HNR, and silence, which approaches the set of acoustic features used by Fandrianto and Eskenazi [5], achieved results similar to the ones reported in their article, in spite of not including dialog-level features. Furthermore, the combination of all acoustic-prosodic features was only able to outperform the acoustic feature set when feature selection was applied, leading to the best results in the subtable, but only with slight improvements.

Finally, by combining the ComParE 2013 set with our acoustic-prosodic features, the result differences in relation to the case when only the first was used were practically negligible, with an improvement of 0.5 percentage points on the Let's Go 2009 dataset and decreases up to 3 percentage points on the remaining datasets. The results obtained by the **Selected** feature set have a similar relation with the ones obtained with the **Selected** set for the ComPaRe 2013 set. However, in this case, the results were only improved on the aggregated dataset, leading to our best result when using all data, with 81.6% accuracy.

Overall, the obtained results show that this task benefits from the use of large feature sets, out of which the most important features can be selected using automatic methods. Furthermore, differential features are important to reduce the effects of speaker dependence. Finally, the information provided by the ComPaRe 2013 feature set seems to be similar to the one provided by our acoustic-prosodic features, since, in general, the combination of both sets did not lead to improved results.

4.2 Unbalanced Datasets

Although we used balanced datasets to train our classifiers, we want them to be able to deal with real situations, where hyperarticulated speech is rare and the focus is on identifying such situations with precision. To assess that capability, we classified all instances of each yearly unbalanced dataset using an RF classifier trained on the aggregated balanced dataset with the **Selected** set of ComPaRe 2013. The first row of Table 2 shows the results obtained by the classifier on each dataset. We can see that, although recall values are high, which means that hyperarticulated speech is identified when it really exists, precision values are low, which means that there are many misclassified examples of non-hyperarticulated speech. This was expected given the lack of balance of the datasets, and we can see that precision decreases as the level of balance decreases. However, these results go against the objective of identifying hyperarticulation with precision. Thus, we looked into the levels of confidence reported by the classifier and performed experiments using different confidence thresholds.

Table 2. **A**ccuracy, **P**recision, **R**ecall, and **F**-measure results obtained on the unbalanced datasets using different **T**hresholds.

	Let's Go 2009				**Let's Go 2012**				**Let's Go 2014**			
T	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**
50%	0.807	0.412	**1.000**	0.584	0.770	0.261	**1.000**	0.414	0.669	0.138	**1.000**	0.242
80%	**0.982**	0.971	0.894	**0.931**	**0.986**	0.920	0.900	**0.910**	0.970	0.649	0.935	0.766
85%	0.965	0.988	0.752	0.854	0.968	0.966	0.633	0.765	**0.985**	0.877	0.831	**0.853**
90%	0.912	**1.000**	0.354	0.523	0.950	**1.000**	0.389	0.560	0.972	**1.000**	0.468	0.638

By analyzing the confidence levels, we noticed that the highest confidence value for an example mistakenly classified as hyperarticulated was 89.6%. In Table 2 we can see that using a fixed threshold of 90% effectively increased precision levels. However, as a trade-off, recall was drastically reduced, which means that many hyperarticulated examples were not identified. We defined two more fixed thresholds – 80% and 85% –. We can see that the threshold leading to the higher F-value tends to increase as the level of balance of the dataset decreases. This hardens the process of selecting a generic threshold. However, we suggest values between 80% and 85%.

Finally, it is important to notice that by performing threshold changes we were able to obtain accuracy results around 98% for every dataset. These values are always above the ones obtained by a chance classifier – 86.5%, 91.9%, and 94.7% for the 2009, 2012, and 2014 datasets, respectively.

5 Conclusions

In this article we presented our approaches on automatic detection of hyperarticulation. We used multiple feature sets and two classification algorithms – SVMs and RFs –, with the latter systematically outperforming the widely used SVMs.

In terms of features, we discovered that this task benefits from large sets of features, out of which the most important can be selected using automatic approaches. This was proved by the results obtained using the ComPaRe 2013 feature set and the combination of all the features we extracted. Furthermore, the feature value differences between the turn being classified and the first in the dialog revealed to be very important features, due to the speaker dependence of many of the extracted features. On the other hand, speech rate and disfluency-based features produced disappointing results, in spite of being classified as relevant features for hyperarticulation detection in the literature.

We achieved accuracy results over 80% on each balanced yearly dataset, as well as on the aggregated dataset. These results surpass the ones obtained by Fandrianto and Eskenazi [5] on similar data by at least 11.6 percentage points.

By modifying the confidence thresholds, we were able to obtain accuracy results around 98% on every unbalanced dataset, while maintaining high precision values.

As future work, we intend to explore features extracted at the phone level, such as the ones described by Soltau and Waibel [18] and Stent et al. [19].

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, by Universidade de Lisboa, and by the EC H2020 project RAGE under grant agreement No 644187.

References

1. Batista, F., Curto, P., Trancoso, I., Abad, A., Ferreira, J., Ribeiro, E., Moniz, H., de Matos, D.M., Ribeiro, R.: SPA: Web-based Platform for Easy Access to Speech Processing Modules. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)
2. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Cortes, C., Vapnik, V.: Support-Vector Networks. In: Machine Learning. pp. 273–297 (1995)

4. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In: Proceedings of the 21st ACM International Conference on Multimedia. pp. 835–838 (2013)
5. Fandrianto, A., Eskenazi, M.: Prosodic Entrainment in an Information-Driven Dialog System. In: Proceedings of INTERSPEECH 2012. pp. 342–345 (2012)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Exploration Newsletter 11(1), 10–18 (2009)
7. Litman, D.J., Hirschberg, J., Swerts, M.: Predicting Automatic Speech Recognition Performance Using Prosodic Cues. In: Proceedings of NAACL. pp. 218–225 (2000)
8. Lopes, J., Chorianopoulou, A., Palogiannidi, E., Moniz, H., Abad, A., Louka, K., Iosif, E., Potamianos, A.: The SpeDial Datasets: Datasets for Spoken Dialogue System Analytics. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)
9. Meinedo, H., Viveiros, M., Neto, J.a.: Evaluation of a Live Broadcast News Subtitling System for Portuguese. In: Proceedings of INTERSPEECH 2008. pp. 508–511 (2008)
10. Moniz, H., Ferreira, J., Batista, F., Trancoso, I.: Disfluency in Spontaneous Speech. In: Proceedings of DISS 2015 (2015)
11. Oviatt, S., MacEachern, M., Levow, G.A.: Predicting Hyperarticulate Speech During Human-Computer Error Resolution. Speech Communication 24(2), 87–110 (1998)
12. Parent, G., Eskenazi, M.: Lexical Entrainment of Real Users in the Let's Go Spoken Dialog System. In: Proceedings of INTERSPEECH 2010. pp. 3018–3021 (2010)
13. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
14. Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M.: Doing Research on a Deployed Spoken Dialogue System: One Year of Lets Go! Experience. In: Proceedings of INTERSPEECH 2006. pp. 65–68 (2006)
15. Ribeiro, E., Batista, F., Trancoso, I., Lopes, J., Ribeiro, R., de Matos, D.M.: Assessing User Expertise in Spoken Dialog System Interactions. In: IberSPEECH 2016 (2016)
16. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K.R., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In: Proceedings of INTERSPEECH 2013. pp. 148–152 (2013)
17. Soltau, H., Waibel, A.: On the Influence of Hyperarticulated Speech on Recognition Performance. In: Proceedings of ICSLP (1998)
18. Soltau, H., Waibel, A.: Acoustic Models for Hyperarticulated Speech. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP. pp. 1779–1782 (2000)
19. Stent, A.J., Huffman, M.K., Brennan, S.E.: Adapting Speaking After Evidence of Misrecognition: Local and Global Hyperarticulation. Speech Communication 50(3), 163–178 (2008)
20. Trancoso, I., Viana, M.d.C., Duarte, I., Matos, G.: Corpus de Diálogo CORAL. In: PROPOR'98 (1998)