AUTOMATIC EVALUATION OF SPOKEN ENGLISH FLUENCY

Om D. Deshmukh, Kundan Kandhway, Ashish Verma

IBM India Research Lab
Vasant Kunj Institutional Area
New Delhi, India
odeshmuk,kkandhwa,vashish@in.ibm.com

*Kartik Audhkhasi**

Dept of Electrical Engineering
Univ of Southern California
Los Angeles, CA, USA
kartik.audhkhasi@usc.edu

ABSTRACT

This paper presents a method to automatically quantify the spoken English fluency skills of speakers. The focus of this work is to automatically compute a numeric score of spoken fluency that is correlated with the numerical score the human assessors would assign. The proposed method combines several novel prosodic and lexical features to compute the fluency score. It is shown that the prosodic and the lexical features provide complementary information for fluency evaluation. Extensive evaluation on human-labeled utterances shows that the proposed technique exhibits similar trends in performance and confusions as shown by human assessors. The proposed technique leads to 84.2% classification accuracy when the two extreme classes of fluency are considered.

Index Terms— fluency evaluation, language learning, prosodic features, lexical features

1. INTRODUCTION

Spoken English fluency is an important skill a call center agent in an offshore facility is expected to possess. The fluency skills are currently evaluated by human assessors. Human assessors typically try to engage the candidates in a conversation for about 10 minutes to evaluate various spoken language parameters. The candidate is asked to speak about a topic in his/her comfort zone (e.g., 'about himself' or 'about his family') in the beginning and the conversation is gradually moved to more spontaneous topics which are outside the candidate's comfort zone. The candidate's fluency skills are evaluated based on his/her participation in the conversation.

It is widely accepted that disfluent speech can be split into three components [1]: the reparandum, the edit phrase and the alteration. The reparandum is the part of the speech signal that the speaker intends to replace. The edit phase is the region between the reparandum and the beginning of the replacement of the reparandum. The edit phrase quite often consists of unfilled or filled pauses (e.g., 'ahh', 'umm') or discourse markers (e.g., 'like', 'you know'). The alteration marks the resumption of fluency. For example, in the sentence: *I helped him ahh her yesterday*; 'him' is the reparandum, 'ahh' is the edit phrase and 'her' is the alteration. The point between the reparandum and the edit phrase is referred to as the Interruption Point (IP).

Many researchers have investigated the problem of detecting and removing disfluencies from speech transcripts. Their main focus was to improve the usability of speech transcripts [2, 3, 4]. Authors in [3] present a transformation-based learning algorithm that uses features based on the word-identity and the Part-of-Speech (POS) tag to learn trends in disfluent speech regions. Shriberg in [5] analyzes the phonetic consequences of disfluencies in spontaneous American English. The insights gained from this analysis are combined with various lexical features to extract metadata information from speech signals [2]. Authors in [4] redefine the speech recognition problem to simultaneously solve other inter-related problems such as identification of POS tags, discourse markers and speech repairs.

The focus of the present work is to automatically quantify the level of spoken English fluency skills of a speaker as perceived by an expert listener and thus differs significantly from the previous works. For example, consider a case where the speaker is structurally fluent but (s)he repeats the same thought several times using similar words. In such cases, the present work is expected to assign a low fluency rating to the speaker as her/his 'fluency of thought' is poor. Automatic evaluation of fluency also involves detecting if the speaker's response is relevant to the topic (s)he is asked to speak on. Note that such a method of fluency evaluation is closely tied to evaluating the vocabulary of the speaker. Our discussions with the expert human assessors confirm this. The ultimate goal of this work is to include a spoken English fluency evaluation module in IBM India Research Lab's spoken language evaluation and learning tool: Sensei [6].

Authors in [7] have developed an automatic spoken fluency scoring technique that computes various features from either the force-aligned output or the free-decoding output of an ASR system. The quantitative assessments of spoken fluency reported in [8] were performed on read speech as against real-time spontaneous speech recordings that are used in the present work. The feature set used in the present work consists of a combination of prosodic features computed directly from the speech signal and lexical features which can be computed from the transcripts (manual or ASR generated) of the speech signal. Many of the features used in the present work are novel and are described in detail in the next section.

2. PROPOSED FEATURES

The proposed fluency evaluation technique uses a combination of prosodic and lexical features to capture the disfluencies in a given speech utterance. The prosodic features are computed directly from the speech signal and are based on detecting extended vowels, filled pauses and regions of silence (i.e., unfilled pauses), all of which are robust indicators of locations of interruption points. Lexical features capture the severity of disfluency. The features are listed in Table 1 and their computation is explained below.

*The work was performed during the summer internship at IBM India Research Lab.

Table 1. *Prosodic and lexical features used for fluency evaluation.*

Prosodic features		
p.a	Average number of filled-pauses per sec.	AvgFP
p.b	Average duration of a filled-pause	DurFP
p.c	Average distance between filled-pauses	DistFP
p.d	Length of the longest filled pause	MaxFP
p.e	Fraction of silence	FracSIL
p.f	Average duration of contiguous silence	DurSIL
p.g	Average duration of contiguous speech	DurSP
p.h	Average distance between silences	DistSIL
Lexical features		
l.a	Count of most frequent word	FreqW
l.b	Total words	TW
l.c	Total unique words	TUW
l.d	Count of filled-pauses	CFP
l.e	Count of dictionary words	Cwrd
l.f	Total repeated 'similar' trigrams	RepTri
l.g	No. of closely occurring unigrams	ClUni
l.h	No. of closely occurring similar trigrams	ClTri

2.1. Prosodic features

While in a conversation, if the speaker senses a delay in either forming the next thought or choosing the set of words to convey the thought, the speaker typically remains silent (i.e., unfilled pause), utters filled pauses (e.g., 'aah', 'umm') or extends the previous syllable (e.g., 'theeeee', 'ammm') to fill the void. In these situations, since the next word is not formulated, the articulators do not change their positions (authors in [9] also make similar observations). As a result, the vocal tract filter characteristics and hence the formants hardly vary over this period. Our algorithm to detect lengthened vocalic regions and filled pauses (jointly referred to as filled pauses, hereafter) is based on this premise.

The algorithm begins by computing the first two formants. The wavesurfer [10] formant-tracker is used in all the experiments reported here. The higher formants were excluded mainly because the first two formants capture the stability adequately and the higher formants are more difficult to track accurately. For a given frame, the standard deviation (SD) of the individual formants in the ± 5 adjacent frames is computed. The SD is typically lower for frames in filled pause as compared to the SD values for frames in normal speech. The distribution of SD values in filled pause frames and in normal speech frames in the training data is used to compute Log Likelihood Ratio (LLR) values. A frame from a test speech utterance is marked as 'probable filled pause' if its corresponding LLR value is positive. Otherwise, it is marked as normal speech frame. Regions with ten or more contiguous 'probable filled-pause' frames are labeled as filled pauses. The above thresholds are optimized on a set of training data with hand-labeled filled pauses. The companion paper [11] describes the above algorithm in greater detail and compares the performance of this algorithm with other standard algorithms for detecting filled pauses.

The prosodic features based on filled pause detection are: (p.a) average number of filled pauses per second (AvgFP), (p.b) average duration of a filled pause (DurFP), (p.c) average distance between consecutive filled pauses (DistFP), and (p.d) length of the longest filled pause. These features are set to zero if no filled pause is detected in a recording. The DistFP feature is set to the duration of the recording (in frames) if the number of filled pauses detected is less than two. The other prosodic features are: (p.e) fraction of silence (FracSIL), (p.f) average duration of contiguous silence (DurSIL), (p.g) average duration of contiguous speech (DurSP), and (p.h) average distance between consecutive silence regions (DistSIL). The silence regions are detected using an energy based Voice Activity Detector (VAD). The energy threshold for the VAD is chosen to minimize the detection of intra-word silences due to stops closures. The utterance-initial and utterance-final silences are discarded from this analysis.

2.2. Lexical features

As a starting point, in the current experiments, the lexical features are computed on manual transcription of the data. The transcripts are processed to remove common stop words and are then passed through the Porter stemmer [12]. Stemming is a process to map various inflected versions of a word to its root form. In this work, stop words are defined as the list of words commonly occuring in the monologues of model speakers (e.g., 'a', 'i', 'is', 'of'). Thus, stop words are not a good indicator of fluency and are removed from the transcripts. The lexical features include: (l.a) count of the most frequent word (FreqW), (l.b) total words (TW), (l.c) total unique words (TUW), (l.d) count of filled pauses (CFP), (l.e) count of dictionary words (Cwrd), (l.f) total repeated 'similar' trigrams (RepTri), (l.g) count of unigrams repeating within a distance of one word (ClUni), and (l.h) total 'similar' trigrams within a distance of two words (ClTri). All these features are normalized by the total duration of the recording to compensate for the variations in the duration of the utterances across speakers. Feature (l.a) captures the speaker's favourite discourse marker, if any. Feature (l.b) is the count of total words in the transcript. This feature is an indirect measure of the proportion of silent regions in the utterance. Feature (l.c) is a good indicator of the speaker's vocabulary. Feature (l.d), which counts the total number of filled pauses in the transcript, is motivated by the observation in [13] that the most common form of disfluency includes a filled pause. Feature (l.e) is the difference of (l.b) and (l.d). In feature (l.f), two trigrams are called 'similar' if they differ by at most a single word deletion, insertion of substitution. For example, consider the following utterance: *I used to go there ahh I go there.* After removing the stop word 'to' and stemming, the utterance becomes: *I us go there ahh I go there*. The trigram *I go there* differs from *I us go there* by only one word and hence the count of 'similar' trigram *I go there* is two. It is rare that the same N-gram is repeated often in a given utterance. Stemming and removing the stop words relaxes the definition of 'repeated N-gram' and incorporating the above definition of 'similar N-gram' relaxes the definition even further thereby increasing the count of repeated N-grams. Features (l.c), (l.e) and (l.f) try to quantify the 'fluency of thought' concept that was alluded to earlier at the end of Section 1. Feature (l.g) captures the number of instances of one-word repetition disfluencies with at the most one-word-long edit phrases while feature (l.h) captures the number of repetition disfluencies where the reparandum is three-word long and the edit-phrase is at the most two-word long.

3. EXPERIMENTS

3.1. Database

The performance of the proposed fluency evaluation technique was evaluated on data collected from real-life assessments of 112 call center candidates. These assessments were conducted at IBM Daksh's call center facility at Gurgaon India. Each of the candidates was asked to first speak about him/herself for about one minute.

Table 2. *Confusion matrix between the two human assessors. The two assessments were done on 72 randomly chosen speakers.*

	1	2	3	4
1	2	6	1	0
2	9	35	9	0
3	3	4	1	0
4	0	1	0	0

The candidate was then asked to speak for about one minute on one of the following topics: (a) favourite movie, (b) favourite vacation, (c) favourite festival, (d) favourite book, or (e) favourite sport. Following this, the candidate was asked to answer a lead question based on the topic (s)he chose. Some of the examples of the lead questions are: 'who is the most favourite character in the movie/book', 'what was the best part of the vacation'. Thus, each speaker records three utterances of about one minute each ('speak-about-self', 'main-topic' and 'lead-question'). The candidate's responses were recorded using a high quality noise-cancellation microphone and stored at a sampling rate of 22050 Hz.

One expert human assessor listened to these responses and rated the spoken English fluency of each candidate on a scale of 1 to 4 where 4 is very fluent, 1 is very disfluent and 2 and 3 are intermediate. 72 of these candidate recordings were also evaluated by a second assessor to estimate the inter-human agreements. Of these 72 recordings, 38 assessments received the same score from both the assessors. Thus, the inter-human agreement is 53.52% $(38/71)$. The confusion matrix between the two assessors is shown in Table 2. Note that of the 33 assessments where the assessors disagree, 29 confusions involve a score of 2 and of the 38 assessments where the two assessors agree, 35 assessments received a score of 2. This indicates that "2" is a very broad class. Indeed, in the larger database of 172 assessments done by the first assessor, 91 candidates received a score of 2 whereas the scores 1, 3 and 4 were received by only 26, 24, and 31 candidates respectively. Of these 91 score-2 assessments, 31 were randomly chosen to be part of the final dataset of 112 candidates.

4. RESULTS

We begin by presenting the Pearson correlation coefficient for some of the features described in Section 2. The correlation coefficient is a good indicator of the linear relationship between the two variables: in our case, the feature and the human score. Disfluent speakers tend to use more and/or longer filled or unfilled pauses than their fluent counterparts. Thus, the AvgFP, DurFP and FracSIL features will have a higher value for disfluent speakers than for fluent speakers resulting in negative correlations with the human fluency scores and DistFP, DurSP and DistSIL will have a lower value for disfluent speakers than for fluent speakers resulting in positive correlations. Among the lexical features, TW and TUW will typically be higher for more fluent speakers and lower for disfluent speakers. Thus, these features are expected to exhibit positive correlation with the fluency scores. The other lexical features, ComW, RepTri, ClUni and ClTri will be lower for fluent speakers and higher for disfluent speakers. Thus, these features should be negatively correlated with the fluency scores. The correlation coefficients for all the prosodic and lexical features are tabulated in Table 3. The second column of the table presents the correlations when the features were computed across all the three utterances for a given speaker. It is evident from the table that the behaviour of these features is in line with our expectations.

Table 3. *Correlation coefficient between the human score and the individual features computed on all the three recordings recorded by the speaker and on the last two recordings.*

Feature	correlation on 3 files	correlation on 2 files
AvgFP	-0.157	-0.228
DurFP	-0.085	-0.173
DistFP	0.003	0.102
MaxFP	0.026	-0.062
FracSIL	-0.177	-0.210
DurSIL	0.011	0.110
DurSP	0.188	0.189
DistSIL	0.119	0.201
FreqW	-0.231	-0.344
TW	0.159	0.143
TUW	0.423	0.494
CFP	-0.205	-0.254
Cwrd	0.268	0.297
RepTri	-0.277	-0.313
ClUni	-0.331	-0.371
ClTri	-0.304	-0.349

In general, the lexical features are more correlated than the prosodic features. The TUW feature, although a very simple feature, is the best single feature in terms of the correlation with the human scores. The decent performance of these relatively simple lexical features raises our hopes that more sophisticated features can lead to a better performance. Some of the more sophisticated features being explored are based on using the POS tag information and on detecting the number of incomplete sentences and fresh starts. On the other hand, to achieve this level of performance from the lexical features in a practical situation the Automatic Speech Recognition (ASR) system has to be highly accurate.

Authors in [14] mention that the rate of repetition disfluency is lower when the speaker has practiced speaking on a topic. We observe similar trends in our data. Almost all the speakers are fluent when talking about themselves: a topic they are familiar with and have potentially rehearsed speaking about. To remove this bias, the same set of features are also computed on only the other two utterances per speaker, namely 'main topic' and 'lead-question'. The corresponding correlation coefficients are tabulated in the third column in Table 3. Note that the correlation coefficients are better for all the features when the 'speak-about-self' utterance is excluded from the analysis. We also conducted experiments using just the 'main-topic' utterance and just the 'lead-question' utterance. Neither case showed any particular improvement over using both the utterances. The rest of the results are presented on features computed on only these two utterances per speaker. The optimal linear combination of the prosodic features leads to a correlation of 0.546 and the optimal linear combination of the lexical features leads to a correlation of 0.598. The optimal linear combination of all the features leads to a correlation of 0.680.

4.1. Classification

The classification experiments reported here, were conducted using the Weka [15] implementation of SVM. The first set of classification results are on the two extreme classes: 1 and 4. As shown in the second column of Table 4, the classification accuracy for this case

Table 4. *Accuracy of the proposed fluency evaluation method.*

	2-class	3-class	4-class
All features	84.21%	71.43%	53.57%
Prosodic	71.93%	55.36%	41.07%
Lexical	78.95%	65.18%	50.89%
Chance	54.38%	49.10%	27.68%

Table 5. *Confusion matrix for 4-class classification*

classified as –>	1	2	3	4
1	14	7	4	1
2	5	14	9	3
3	5	9	9	1
4	1	6	1	23

is 84.21% when all the features are used. The corresponding chance accuracy is tabulated in the last row for comparison. Chance accuracy is defined as the accuracy when all the test samples are classified as belonging to the class with majority samples. It is evident that the proposed technique can reliably discriminate between the best and the worst candidates. The third and the fourth rows compare the corresponding classification accuracies when only prosodic and only lexical features are used respectively. The performance of the lexical feature-set is better than that of the prosodic feature-set although neither of them is close to the performance of the combined feature set. This indicates that the prosodic and the lexical features provide complementary information for fluency evaluation.

The next set of classification results are on three classes, where '2' and '3' are combined to form one class and 1 and 4 are the other two classes. The third column of Table 4 presents the accuracy when all the features are used and when the prosodic and lexical feature-sets are used separately. Finally, the fourth column presents the results on 4-class classification. As expected, the performance drops as the number of classes is increased although the accuracy in each case is much higher than the chance accuracy. Our analysis indicates that one of the main reasons for the drop in performance when utterances with intermediate scores are included is that often the human assessors are very accurate in identifying candidates who fall in either of the two extremes but assign intermediate scores, especially the score '2', to candidates with a broad range of fluency skills. The confusion matrix for the 4-class classification using all the features is tabulated in Table 5. Note that the maximum confusion is between class 2 and class 3 whereas class 1 and class 4 are quite well separated. This trend is similar to the trend observed in confusions among the human assessors (ref. Table 2).

5. DISCUSSION AND FUTURE WORK

In this paper, we propose an automatic spoken fluency evaluation technique to match the perception of expert human assessors. The proposed technique uses a combination of novel prosodic and lexical features to compute an overall score of fluency. The prosodic and the lexical features contribute complementary information for fluency evaluation. It is shown that the proposed technique and the human assessors make similar types of errors. Work is in progress to compute the lexical features from the output of an ASR system and to analyze the effect of ASR's word error rate on the overall fluency evaluation.

6. ACKNOWLEDGEMENTS

Authors would like to thank Shajith Iqbal, colleague from IBM India Research Lab and the assessors from IBM Daksh for fruitful discussions.

7. REFERENCES

[1] W. J. M. Levelt, "Monitoring and self-repair in speech," in *Cognition*, 1983, pp. 41–104.

[2] Y. Liu et. al., "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 175–187, June 1992.

[3] M. Snover et. al., "A lexically-driven algorithm for disfluency detection," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, Boston, 2004, pp. 157–160.

[4] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue," *Computational Lingustics*, vol. 25(4), pp. 527–571, 1999.

[5] E. Shriberg, "Phonetic consequences of speech disfluency," in *Int. Conf. on Phonetics Sciences*, San Francisco, 1999, pp. 619–622.

[6] A. Chandel et. al., "Sensei: Spoken language assessment for call center agents," in *IEEE Intl Workshop on ASRU*, Kyoto, Japan, December 2007.

[7] K. Zechner and I. Bejar, "Towards automatic scoring of non-native spontaneous speech," in *Proceedings of the HLT*, New York, 2006, pp. 216–223.

[8] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *Journal of Acoust. Soc. of Am.*, vol. 107(2), no. 2, pp. 989–999, Feb. 2000.

[9] Masataka Goto et. al., "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. of Eurospeech*, September 1999, pp. 227–230.

[10] Wavesurfer: An open source speech tool, "http://www.speech.kth.se/wavesurfer/," .

[11] K. Audhkhasi et. al., "Automatic detection of filled pauses for fluency evaluation," in *submitted to Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009.

[12] M. F. Porter, "An algorithm for suffix stripping," in *Program*, 1980, pp. 130–137.

[13] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on rcognized words," *Proc. ICSLP*, pp. 2247–2250, 1998.

[14] H. Branigan et. al., "Non-lingustic influences on rates of disfluency in spontaneous speech," in *Int. Conf. on Phonetics Sciences*, San Francisco, 1999.

[15] S. Garner, "Weka: The waikato environment for knowledge analysis," in *Proc. of New Zealand Computer Science Research Students Conference*, 1995, pp. 57–64.