



Role of language models in spoken fluency evaluation

Om D Deshmukh, Harish Doddala, Ashish Verma, Karthik Visweswariah

IBM Research India, Vasant Kunj Institutional Area, New Delhi, India

{odeshmuk, harish.dh, vashish, v-karthik}@in.ibm.com

Abstract

This paper addresses the task of automatic evaluation of spoken fluency skills of a speaker. Specifically, the paper evaluates the role of language models built from fluent and disfluent data in quantifying the fluency of a spoken monologue. We show that features based on relative perplexities of the fluent and the disfluent language models on a given utterance are indicative of the level of spoken fluency of the utterance. The proposed features lead to a spoken fluency classification accuracy of 39.8% for 4-class and 68.4% for 2-class classification. Combining these features with a set of prosodic features leads to further improvement in the classification accuracy thus highlighting the complementarity of the information they contribute compared to the low-level disfluency information captured by the prosodic features.

Index Terms: spoken fluency evaluation, language models, perplexity

1. Introduction

Evaluating spoken fluency is an important component in evaluating the overall spoken language skills of a candidate. The sustained interest of a wide population in acquiring spoken English skills and the success and demands of the off-shore call centers have contributed towards active research in the general area of computer assisted language learning and also specifically in the area of automatic spoken fluency evaluation.

In a typical spoken fluency evaluation setup, the candidate is asked to speak for a few minutes on an extempore yet familiar topic(s) (for example, 'favourite movie', 'favourite vacation'). The candidate's response is then analyzed to compute a numeric fluency score. The analysis can include computation of features directly from the speech signal, called the prosodic features, or computation of features from the manual or Automatic Speech Recognition (ASR)-generated transcripts of the speech signal, called the lexical features, or a combination of the two.

The fluency evaluation technique proposed in [1] uses lexical features which are derived from either force-alignment or full decoding of spontaneous speech using a standard ASR system. A total of 23 lexical features are used to capture content richness, vocabulary, rate of insertion of filled-pauses and rate of speech. On a 5-point scale, the classification accuracy is 47% as compared to the assign-majority-class accuracy of 25%. In [2], spoken fluency is evaluated using the Rate-of-Speech (RoS) feature, defined as the ratio of the number of phones in the speech signal and the total duration of the speech signal. This feature is computed on the ASR output of spontaneous, read and repeated speech utterances. On a 5-point scale, the correlation between the RoS feature and human-assigned score varies between 0.5 and 0.6. Authors in [3, 4] analyze the correlation between human spoken fluency scores and various

temporal-rate based features computed on the ASR output of read and spontaneous speech of native and non-native Dutch speakers. The study establishes that the correlation between the human scores and the features is higher for read speech than for spontaneous speech. In our earlier work [5, 6], we proposed a combination of filled-pause and silence based prosodic features and lexical features based on manual transcripts of spontaneous speech for automatic classification of spoken fluency. It was shown that using 16 features, the classification accuracy is 53.8% for 4-class, 71.4% for 3-class and 84.2% for 2-class classification. It is worth mentioning that the accuracies are high largely because the lexical features were computed on manual transcripts as opposed to being computed on (error-prone) ASR transcripts. In [7] authors have developed techniques based on various prosodic and/or lexical features to detect disfluencies, but the focus in those methods is to purge the speech transcripts of the disfluencies to improve the accuracy and efficiency of down-stream text processing.

In this paper, we explore an approach to evaluate the fluency of an utterance based on the relative perplexities of Language Models (LMs) trained on different domains with respect to this utterance. This work presents a significant shift in the role the LMs play in typical ASR-related applications. In a typical ASR setup, while recognizing a test utterance, the LM suggests the set of most likely next words given the previous N words (N=2 for trigram LMs). The ASR system then compares the acoustic evidence from the current and the next speech frames with the acoustic models corresponding to only these most-likely words. The relevance of a given LM for a test domain is measured by computing the perplexity on a sample test data set which is representative of the test domain. Perplexity is an information theoretic way of capturing the ability of a LM in predicting the word sequences in a given text utterance. Formally, perplexity (PP) is defined as [8]:

$$PP = P(w_1, w_2, \dots, w_K)^{-\frac{1}{K}}$$

where $P(w_1, w_2, \dots, w_K)$ is the probability the LM would assign to observing the word sequence w_1, w_2, \dots, w_K . For a trigram LM, this probability can be estimated as:

$$P(w_1, w_2, \dots, w_K) = \prod_{i=1}^K P(w_i | w_{i-2}, w_{i-1})$$

where $P(w_1)$ and $P(w_2 | w_1)$ are the backoff unigram and bigram probabilities, respectively. Perplexity can be considered to be a measure of the number of words that are equally likely to follow a given sequence of words on an average.

While LM perplexity is typically used to measure how well a language model can predict a given domain, the current work derives features based on the LM perplexity measure to estimate how closely aligned the given utterance is to the domain represented by the LM. The domains modeled by the two LMs are 'fluent' and 'disfluent' domains.

The rest of the paper is organized as follows: Section 2 describes the various LMs used in this study and the choice of LM training data. Section 3 describes the proposed features based on relative perplexities that can be used to quantify fluency. The prosodic features are briefly discussed in Section 4. The various comparative experiments and the corresponding classification results are presented in Section 5. Section 6 summarizes the findings of the experiments and outlines the ongoing and future direction of research in this area.

2. Language Models for Fluent and Disfluent Domains

In ASR applications, the LM is trained on text data collected from the domain on which the ASR will be applied. For example, if the ASR is to be applied for recognition of read news, the LM is trained on news articles collected from leading news portals such as BBC News, CNN, Times of India and so on. On the other hand, if the ASR is to be applied for recognition of a conversation between a customer and an agent at a technical support call center, the LM is trained on FAQs and manuals of the technical product and transcripts of sample conversations at the call center. In this work, instead of treating the extempore topics as the different domains for LM, we treat fluent speech and disfluent speech as the two domains on which the LMs should be trained. The challenge for such an approach is to obtain substantial disfluent data which also has its corresponding fluent counterpart. Manual transcripts of the Switchboard database [9] are well suited for this purpose. The Switchboard database consists of spontaneous and natural conversations between speakers on a variety of everyday topics (e.g., public education, gardening, consumer goods and so on). The database contains about 2430 conversations spoken by 543 speakers comprising about 3 million words of text. The database was designed mainly to support research activities in telephony speech recognition. The standard transcription of the database includes word fragments (e.g., wh-, you kn-).

As part of the Penn treebank project [10], various disfluencies in the switchboard data were annotated in the transcripts. These annotations include non-sentence elements such as fillers (e.g., oh,uh,um), explicit editing terms (e.g., 'I mean', 'excuse me'), discourse markers (e.g., 'like', 'you know'), simple and complex restarts with and without repairs. The example given below illustrates the various disfluency annotations:

{C and, } {F uh, } {D you know, } we
get [c-, + cars] on our block, {F uh, }
regularly, {F uh, } {F uh, } [g-, + F uh,
gone] through, rifled through and stuff

'{C and}' indicates that 'and' is a coordinating-conjunction; '{F uh}' indicates that 'uh' is a filler; '{D you know}' indicates that 'you know' is a discourse marker; 'g-' indicates a word fragment; '[g-, + {F uh, } gone]' indicates a restart which has a filler nested within it. Further details on the disfluency annotation details can be found in [11]. The disfluency annotations for 1155 conversations comprising 205,000 utterances and 1.4 million words are available as part of the Switchboard Dialog Act Corpus [9]. In the current work, these annotations are used to train the disfluent and fluent LMs. The disfluent version of an annotated sentence, as the one shown above, is obtained by removing the annotation meta-data and retaining all the spoken words. To obtain the fluent version of the annotated sentence, (a) the coordinating-conjunctions, fillers, discourse markers and word fragments are removed, (b)

Table 1: *Perplexities of the disfluent, fluent and combined language models on the data used to train the disfluent and the fluent language models*

	fluent-data	disfluent-data
fluent-LM	15.82	24.04
disfluent-LM	18.83	16.81
combined-LM	26.58	26.78

in restarts with repairs only the final repair is retained, and (c) in restarts without repairs the restarts are deleted. The disfluent and fluent versions of the example utterance mentioned above are:

Disfluent version: and uh you know we get c-cars on our block uh regularly uh uh g- uh gone through rifled through and stuff

Fluent version: we get cars on our block regularly gone through rifled through and stuff

It is worth mentioning that the annotation also identifies utterances that are incomplete in one turn but are completed across multiple consecutive turns. Our data preparation process combines these incomplete chunks across the turns to form one complete utterance for both disfluent as well as fluent versions. Turns with less than three words (most of these are labeled 'backchannel') are not included in the LM training data as they would bias the LMs towards short sentences which are unlikely in the current fluency evaluation setup where the speaker is expected to talk continuously on a topic. The data processing steps lead to 82,184 utterances with 1.24 million words and 18,716 unique words for the disfluent version and 75,595 utterances with 0.85 million words and 17,622 unique words for the fluent version.

In the current fluency evaluation setup the candidates have an Indian background and typically talk about Indian movie stars, Indian sports stars or Indian vacation spots whereas the Switchboard data has content that typically refers to American activities. To reduce this mismatch, 28344 sentences (512645 words with 18405 unique words) from various Indian news portals were added to both disfluent and fluent versions of the Switchboard data. The two LMs are standard trigram LMs trained using IBM's SLM toolkit with modified Kneser-Ney smoothing.

To establish that the two LMs have indeed captured the differences in the sentence structures of disfluent and fluent sentences, perplexities of the two LMs are computed on the data used to train each of the two LMs. The perplexities are tabulated in Table 1. As expected, the perplexity of the fluent LM on the fluent data is lower than its perplexity on the disfluent data. Similarly, the perplexity of the disfluent LM on the disfluent data is lower than its perplexity on the fluent data.

It can also be inferred from the table that it is more difficult for the fluent LM to predict disfluent sentences than for the disfluent LM to predict fluent sentences. This is to be expected because the disfluent LM training data contains several N-grams which are also present in the fluent LM training whereas the disfluent LM training data has several N-grams which are not present in the fluent data.

A third LM, called the combined-LM, is trained on the combination of the data used to train the disfluent and the fluent LM. The combined-LM is not biased towards either fluent utterances or disfluent utterances as can be seen from the final row in table

1 which compares the perplexity of the combined-LM on the fluent and disfluent training data. As described in Section 5.2, the candidate recordings are decoded using the combined-LM as decoding these recordings with either the fluent-LM or the disfluent-LM will unfairly bias the ASR-output towards either fluency or disfluency, respectively.

3. Perplexity-based feature computation

As mentioned earlier, perplexity of a LM on a test utterance is indicative of the suitability of the LM in predicting the word sequences of the utterances in the test domain. Thus, the perplexity computation can be used to estimate the probability of observing each of the words in the utterance given the word history. Note that this sequence of words also includes the 'sentence-begin' and 'sentence-end' tokens. Thus, the probability of observing a given word in the utterance will depend not only on the word history in the utterance but also on the location of the sentence boundaries. For example, consider the utterance *i am in the middle of a sentence*. The perplexity of the fluent-LM on this utterance is 15.92. Now, consider the same utterance but with a sentence boundary as shown below: *i am in the . middle of a sentence*. The perplexity of the fluent-LM on this 2-sentence utterance is 43.97! Clearly, adding artificial sentence boundaries adversely affects the probability computations and hence the perplexity measure.

The ASR output does not contain any punctuation marks including the sentence boundary indicator: full-stop. Current work uses a silence-duration based heuristic to detect the sentence boundaries in the ASR output. A sentence boundary is said to have occurred if the duration of the consecutive silence frames as detected by the ASR is above a threshold of T ms. We experimented with various values of T on a few sample ASR outputs and found that varying T over a range of [250 – 750] ms leads to reasonable sentence boundaries. In the present work, T is set to 250.

Given the sentence-delimited output of the ASR system on a test utterance, the proposed relative perplexity based features, are computed as follows: As a first step, the LM-prediction probabilities on words which occurred infrequently in both the fluent-LM and the disfluent-LM training datasets are excluded. The LM-prediction probability for infrequent words is quite low and is more dependent on the size of the LM training data than on the word history context. A word is marked as infrequent if it occurs less than 50 times in both the training datasets. The following four features are then derived as follows:

- Number of fluent-unlikely words (NFU): NFU is defined as the proportion of the decoded words which have a low prediction probability with respect to the fluent LM. The fluent-LM prediction probability could be low because of one of the following two reasons: (a) the context is truly disfluent, or (b) the context is grammatically incorrect either due to high ASR-error or incorrect sentence structure formation by the speaker. The threshold for 'low prediction probability' is set to 0.0001.
- Number of strong-fluent-unlikely words (NSFU): NSFU is defined as the proportion of the decoded words which have a low prediction probability with respect to the fluent-LM but a relatively high prediction probability with respect to the disfluent-LM. In the current work, the high prediction probability threshold is set to 0.001. These words should ideally be the locations of disflu-

ency in the speech utterances. Indeed, the correlation of this feature with the human score is 0.214 which is the highest among other perplexity-based features.

- Reliable relative perplexity (RRP): RRP is defined as the difference in the perplexities of the two LMs on the test utterance. The prediction probabilities of infrequent words are excluded for this computation.
- Relative characteristics words (RCW): Comparing the training data used for the two LMs showed that there are certain words which occur significantly more number of times in one of the two LMs. For example, the word 'know' occurs 16168 times more in the disfluent training data than in the fluent training data. This motivated the RCW feature which captures the relative perplexity measures for only these words. In the current work, a word is defined as a 'characteristic' word if the difference in the number of occurrences of the word in the two training sets is more than 50.
- Global relative perplexity (GRP): GRP is defined as the difference in the perplexities of the two LMs computed on the entire test utterances. Experiments using the GRP feature show that this feature, by itself, is not discriminative of fluent vs. disfluent utterances.

4. Prosodic Features

The performance of the proposed relative perplexity feature in evaluating spoken fluency is compared with that of a set of prosodic features.

The prosodic features used in this study are based on detecting filled pauses and silence regions. Both of these are good indicators of disfluency as hesitation in thought formation or choice of words typically leads the speaker to either utter filled pauses ('ahh', 'umm') or remain silent. Filled pauses are identified by detecting regions of high formant stability. Since the vocal tract remains steady during the production of filled pauses, the vocal tract filter characteristics and hence the filter resonances (i.e., the formants) remain steady during the production of filled pauses. We have shown that the formant-stability-based filled-pause detection algorithm performance superior to other standard filled-pause detection techniques [5]. The silence regions in the speech signal are detected by using an energy based Voice Activity Detector (VAD). The final prosodic features are: (a) average number of filled pauses per second, (b) average duration of a filled pause, (c) average distance between consecutive filled pauses, (d) length of the longest filled pause, (e) fraction of silence in the speech signal, (f) average duration of contiguous silence, and (g) average distance between consecutive silence regions. For further details please refer to [6].

5. Experiments

5.1. Database

The performance of the proposed features on spoken fluency classification is 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 candidate was asked to first speak about him/herself for about one minute. 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.

Table 2: Accuracy of the various features on spoken fluency evaluation.

	2-class	3-class	4-class
baseline	54.4%	49.1%	27.7%
Prosodic features	66.8%	50.9%	33.1%
proposed features	68.4%	51.9%	39.8%
proposed+prosodic	73.7%	54.5%	41.1%

An example of a lead questions is: 'who is the most favourite character in the movie/book'. Thus, each speaker records three utterances of about one minute each. 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 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. The human score is the ground truth used to evaluate the performance of the proposed technique. A subset of this data was evaluated by one more human assessor. The inter-assessor agreement was only 53.5% which highlights the subjectivity in human evaluation and the complexity of the task.

5.2. ASR

The ASR system used in the present work is the research version of IBM's large vocabulary continuous speech recognition system [12]. The front-end consists of 24-dimensional MFCCs computed at a frame rate of 10ms. For each frame, the MFCCs from the neighbouring ± 4 frames are grouped together. Linear Discriminant Analysis (LDA) is then used to transform these 24 X 9 dimensional feature vectors into 60-dimension vectors. The acoustic models used by the ASR system are context-dependent models trained on about 130 hours of speech data consisting of more than 500 Indian English speakers. As mentioned earlier, the ASR system uses the combined-LM which is trained on the combination of the data used to train the disfluent and the fluent LM. Such a LM ensures that the decoding is not unfairly biased towards either fluency of disfluency.

5.3. Results

The spoken fluency classification experiments reported here, were conducted using the Weka [13] implementation of SVM. Table 2 compares the classification accuracy of the proposed features with that of the prosodic features and also when the two sets of features are combined. Baseline accuracy, defined as the accuracy when all the test samples are classified as belonging to the class with majority samples, is also tabulated. Classification results are presented for the 2-class classification case where only the two extreme classes are considered, for 3-class classification where the two 'intermediate fluency' scores of 2 and 3 are combined and for 4-class classification.

The accuracy numbers for prosodic features are slightly different than the ones presented in [6] mainly because of the differences in the implementation of the algorithms for prosodic features computation. The performance of the proposed features is consistently better than that of the prosodic features and combining the two sets of features improves the accuracy.

6. Discussion and Future Work

This paper explores the use of two LMs trained on fluent and disfluent data for spoken fluency evaluation. Features based on the relative prediction probabilities of the decoded words in the utterances by the two LMs are used for fluency classification.

The performance of the proposed perplexity-based features can be further improved by a more appropriate choice of data for training the fluent and the disfluent LMs. The perplexity of the combined-LM on the manual transcripts of the test database is 335.6 when the prediction of unknown words is excluded from perplexity computation. The perplexity jumps to 581.6 when the prediction of unknown words is included! The manual transcripts have a total of 69629 words of which 6354 are unknown words for the combined-LM. Thus, the rate of out-of-vocabulary words is close to 9%. This also translates to a relatively high WER of about 50%. In spite of the high perplexity, the switchboard data was used in this work mainly because of the easy availability of disfluent and corresponding fluent transcripts. Current efforts are focused on learning disfluency patterns from the switchboard data to generate disfluent data for a new domain for which fluent data is available.

Current efforts are also focused on computing lexical features on the ASR output and proposing refinements to reduce the drop in performance, if any, of the lexical features due to the ASR errors.

7. References

- [1] K. Zechner and I. Bejar, "Towards automatic scoring of non-native spontaneous speech," in *Proceedings of the HLT*, New York, 2006, pp. 216–223.
- [2] C. V. D. Walt F D Wet and T. Niesler, "Automatic large-scale oral language proficiency assessment," in *Proceedings of the Interspeech*, Antwerp, Belgium, 2007, pp. 218–221.
- [3] C. Cucchiari, 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.
- [4] C. Cucchiari and H. Strik, "Automatic assessment of second language learners' fluency," in *Proceedings of the ICPhS*, San Francisco, USA, 1999, pp. 759–762.
- [5] K. Audhkhasi et. al., "Automatic detection of filled pauses for fluency evaluation," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009.
- [6] O. Deshmukh et. al., "Automatic evaluation of spoken english fluency," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009.
- [7] 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 2006.
- [8] S. Young et. al., "The htk book," Tech. Rep., Cambridge University Engineering Department, 2006.
- [9] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," Tech. Rep., Linguistic Data Consortium, Philadelphia, 1997.
- [10] M. Marcus et. al., "Building a large annotated corpus of english: the penn treebank," *Computational Linguistics*, vol. 19, 1993.
- [11] M. Meteer et. al., "Dysfluency annotation stylebook for the switchboard corpus," Tech. Rep., Linguistic Data Consortium, Philadelphia, 1995.
- [12] B. Ramabhadran, O. Siohan, and A. Sethy, "The IBM 2007 speech transcription system for european parliamentary speeches," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Japan, December 2007, pp. 472–477.
- [13] S. Garner, "Weka: The waikato environment for knowledge analysis," in *Proc. of New Zealand Computer Science Research Students Conference*, 1995, pp. 57–64.