Lattice-based Lexical Cues for Word Fragment Detection in Conversational Speech

Kartik Audhkhasi, Panayiotis Georgiou, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL)
Electrical Engineering Department
University of Southern California
Los Angeles, CA 90089-2564
audhkhas@usc.edu, {georgiou,shri}@sipi.usc.edu

***Abstract*—Previous approaches to the problem of word fragment detection in speech have focussed primarily on acoustic-prosodic features [1], [2]. This paper proposes that the output of a continuous Automatic Speech Recognition (ASR) system can also be used to derive robust lexical features for the task. We hypothesize that the confusion in the word lattice generated by the ASR system can be exploited for detecting word fragments. Two sets of lexical features are proposed - one which is based on the word confusion, and the other based on the pronunciation confusion between the word hypotheses in the lattice. Classification experiments with a Support Vector Machine (SVM) classifier show that these lexical features perform better than the previously proposed acoustic-prosodic features by around 5.20% (relative) on a corpus chosen from the DARPA Transtac Iraqi-English (San Diego) corpus [3]. A combination of both these feature sets improves the word fragment detection accuracy by 11.50% relative to using just the acoustic-prosodic features.**

I. Introduction

Unrehearsed, conversational speech, is typically characterized by a notable presence of disfluencies such as word fragments, filled pauses, hesitations and repeats. Levelt [4] proposed what is now called the Repair Interval Model (RIM) of disfluent speech. According to this model, disfluent speech consists of three phases: reparandum, edit phase and alteration. Reparandum refers to the region of the speech signal that the speaker intends to replace. The edit phase denotes the region between reparandum and alteration, and is usually characterized by filled pauses (e.g. UM, AH etc.) or silence. The alteration region signifies the resumption of fluency. The point between the reparandum and the edit phase where the speaker departs from fluency is called the interruption point. Detection of disfluencies in speech has been the subject of quite a few works in the past [5], [6], [7], [8], [9].

The focus of this paper is on the detection of word fragments. Accurate and robust detection of word fragments is important from many viewpoints. First, they constitute an appreciable fraction of all disfluencies. As reported in [1], around 17% of the disfluencies in the Switchboard corpus [10] are word fragments. The knowledge of their location can hence be used as part of disfluent speech detection. Second, since word fragments are not explicitly part of the vocabulary of an ASR system, they are invariably misrecognized as some other word in the vocabulary. This significantly increases the Word Error Rate (WER) of the system. Equipped with the knowledge of the location of these word fragments, one can annotate the ASR output for their presence, thus enabling better readability and further processing of the output text.

One potential area of application could be in a Speech to Speech (S2S) translation system. Modern day S2S systems [11] are mainly composed of an ASR, a Statistical Machine Translation (SMT) unit and a Text To Speech (TTS) synthesis system, organized in a pipeline. The disfluencies need to be identified at the output of the ASR stage. Having such a facility could arguably improve the performance of the SMT and the overall S2S system in general. This will be one of the focus areas of our subsequent work.

This paper is organized as follows. Section II gives an overview of the acoustic-prosodic features reported in [1]. These features serve as a baseline for the proposed work. In section III, we discuss the proposed word lattice-based lexical features. Section IV explains the experimental setup and provides the classification results. We conclude the paper in Section V and provide directions for future work.

II. Acoustic-Prosodic Features

The acoustic-prosodic characteristics of speech disfluency have been well explained in [12]. The features proposed by Liu in [1] utilize some of these attributes for the purpose of detecting word fragments. These acoustic-prosodic features belong to two broad categories: prosodic features and voice quality measures. They are discussed in the following subsections.

A. Prosodic Features

The first set of prosodic features are based on the fundamental frequency (F0) of speech. These include the change in the average F0 of a word as compared to the speaker's average value and the change in F0 value across a word boundary. A third F0-based feature is the log ratio between the minimum F0 before the word boundary and the maximum

value after the boundary. The second set of features is based on the frame energy of the speech signal, and is computed in a way similar to the F0-based features (excluding the log ratio feature). The final set of prosodic features is based on duration. This includes word duration, pause duration and duration of the last phoneme of a word. These acoustic-prosodic features are extracted after forced alignment of the speech transcript with the audio using a trained ASR system.

B. Voice Quality Measures

Liu [1] proposes that in addition to the above prosodic features, some measures which capture the quality of voice are also important for the detection of word fragments. The first of these measures is jitter, which quantitatively captures the perturbation in the pitch period of the speaker. It has been previously used to identify pathological speech [13], and is defined in (1), where N denotes the number of intervals in the point process and T_i is the time associated with the i^{th} point. This point process is generated by detecting the occurrence of amplitude peaks of the glottal pulse train.

$$J = \frac{\sum_{i=2}^{N-1} |2T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i} \tag{1}$$

The second set of voice quality measures is computed from the spectral tilt, which is defined as the slope of the least squares line fit to the spectral envelope of the speech frame. It is measured in dB/octave, and has been shown to be a good indicator of syllable stress and breathiness of the voice [1]. The maximum, minimum and mean of the spectral tilt over all frames in a given word are used as voice quality features for word fragment detection.

The final set of features is related to the Open Quotient (OQ), which is defined as the fraction of the time in a glottal cycle when the vocal folds are open. OQ has been shown to be a good measure of the breathiness and creakiness of speech [14]. Fant [15] proposed an approximation to the OQ in the spectral domain as given in (2); H_1 and H_2 are the amplitudes of the first and second harmonics of the speech spectrum. Liu [1] approximates them by $F0$ and $2 * F0$ respectively, where $F0$ is the fundamental frequency. The maximum, minimum and mean OQ over all voiced frames in a word are used in the final voice quality features.

$$OQ = \frac{1}{5.5}\Big\{ \log\Big(\frac{H_1 - H_2 + 6}{0.27}\Big)\Big\} \tag{2}$$

III. Word Lattice-based Lexical Features

The acoustic-prosodic features proposed by Liu [1] aim to capture the distinct characteristics in the speech signal near a word fragment. Since a word fragment is not part of fluent speech, it is rightly expected that various characteristics of the speech signal such as the fundamental frequency (F0), energy and jitter will deviate from their nominal values found in fluent speech. However, there are additional features that can be advantageously used to model word fragments. By their very nature, word fragments can be viewed as "Out of Vocabulary" (OOV) words in the context of conversational speech recognition, and thus can be considered to distort the underlying structure of the language as well. This distortion can be captured by an ASR system, which not only takes the acoustics into account, but also the structure of the language by means of a language model. Hence, we hypothesize that the output of an ASR system can help derive useful lexical features to detect these word fragments. In particular, the idea is to capture the degree of confusion of the ASR system at every time instant (analysis frame).

The word lattice is a suitable form of ASR output for this purpose. It is a directed acyclic graph which contains the set of all possible word hypotheses corresponding to a given (decoded) speech signal. The lattice consists of nodes corresponding to the various hypothesized word boundaries, and links containing the words with their acoustic and language model scores. For computing the proposed lexical features, word lattices have an advantage over N-best lists, since they capture a much greater number of possible word hypotheses corresponding to every time instant in the utterance. Two types of lattice-based confusion features are proposed: word confusion-based and pronunciation confusion-based. They are described in the following subsections.

A. Word Confusion-based Features

Entropy provides a simple measure of the confusion between various word hypotheses at a given time instant in the lattice. It is computed as given in (3), where M is the number of unique word hypotheses in the lattice at the time instant of interest and $\hat{p}(i)$ is the relative frequency of the i^{th} word.

$$Entropy = \sum_{i=1}^{M} \hat{p}(i) \log_2 \frac{1}{\hat{p}(i)} \tag{3}$$

Since a word fragment is an anomaly for the ASR system both from the perspective of acoustics and language, it is expected that the entropy of word hypotheses will be higher during word fragments as compared to normal (non-fragmented) words. This can be seen in Fig. 1, where roughly 65.4% of the non-fragmented words have a mean entropy of less than 4, whereas around 69.2% of the word fragments have a mean entropy above 4. It must be noted however that a high Word Error Rate (WER) of the ASR system can reduce the discrimination between the two classes based on entropy. This is attributed to the general increase in the word confusion in the lattice at all time instants, including for non-fragmented words. Fig. 1 uses lattices from an English ASR system having a WER of 40% on the classification corpus described in section IV.A. One can expect the histogram for normal (full or non-fragment) words to peak a lot more for an ASR system with a lower WER. In addition to the mean, the minimum and the maximum entropy of the hypotheses over all time instants (frames) of a word are also used as features.

Two additional features which capture the word confusion are the number of unique word hypotheses and the number of occurrences per word hypothesis. Time instants during a word

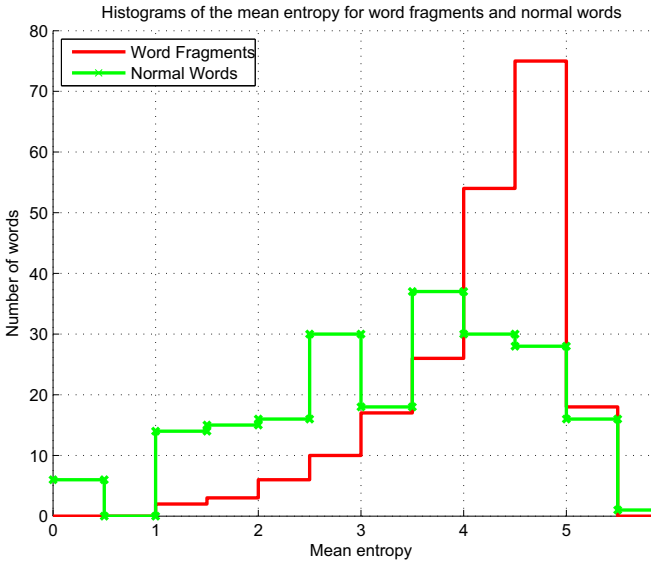

Fig. 1. Histograms of the mean entropy across the ASR output lattice of the word hypotheses during word fragments and normal words. 211 samples each of word fragments and normal words were taken from the San Diego corpus [3].

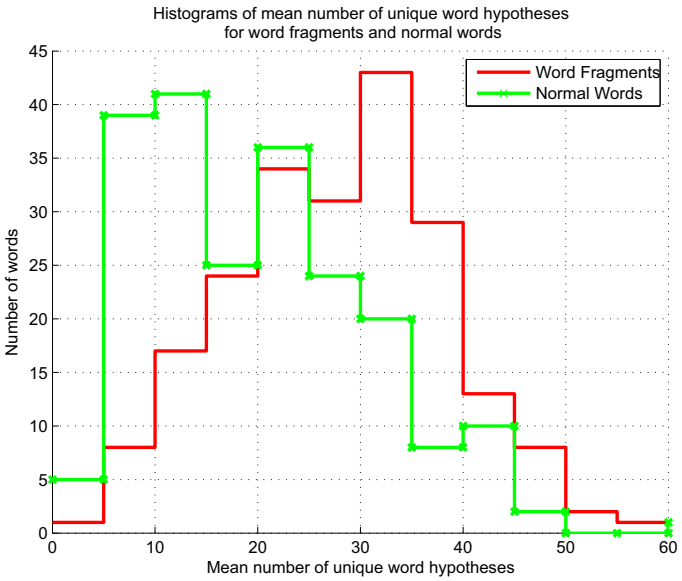

Fig. 2. Histograms of the mean number of unique word hypotheses across the ASR output lattice during word fragments and normal words. 211 samples each of word fragments and normal words were taken from the San Diego corpus [3].

fragment are expected to show a greater number of unique word hypotheses, and fewer occurrences per word. Fig. 2 shows the histograms for the mean number of unique word hypotheses for the two classes. In our experiments, around 69.2% of the full words have a mean number of unique word hypotheses fewer than 25. On the other hand, roughly 60% of the word fragments have more than 25 mean unique word hypotheses.

B. Pronunciation Confusion-based Features

In addition to the statistics of the word hypotheses, it is also necessary to take into account their pronunciation similarity. For example, a specific frame in a lattice contains the word hypotheses `ACADEMIA, ACADEMIC, ACADEMICALLY, ACADEMICS, ACADEMIES` with counts 15, 30, 30, 9 and 16. The entropy for this frame will be 2.0125. But, we can see that all these words have pronunciations that are very close to one another, and thus signify less confusion as compared to (for example) the following set of words: `DEMOLISH, DIATRIBE, ADMONISH, DIVIDE, TIME` with identical counts as the previous example. Hence, a measure of pronunciation similarity between the various word hypotheses can serve as a useful feature for word fragment detection.

We propose the use of Minimum Edit Distance (MED) or Levenshtein distance [16] between the phonetic baseforms of all pairs of word hypotheses as a measure of this similarity. The MED between two strings is defined as the minimum number of operations needed to transform one string into another, where an operation could include an insertion, deletion or substitution. The MED between two strings can be computed using a bottom-up dynamic programming algorithm [16]. We use a value of 1 for the insertion, deletion and substitution penalties during experiments.

As mentioned earlier, we compute the MED between the baseforms of the two word hypotheses from a dictionary. At every time instant, the average MED between all possible pairs of word hypotheses is computed. Since there are invariably multiple occurrences of each word hypotheses, the average MED at this time instant can be thought of as a weighted sum of the MED between all possible pairs of unique word hypothesis. As in the case of other lexical features, the minimum, maximum and average of this weighted MED is computed over all the time instants (frames) in the word and used in the final feature set. Fig. 3 shows the histograms of minimum average MED over all frames belonging to word fragments and normal words. Only 19.4% of the word fragments have a minimum average MED less than 2, although the percentage of such full words is around 40%. One can expect an even higher percentage of normal words with minimum average MED below such a threshold for an ASR with a low WER.

IV. Experiments and Results

For generating the word lattices, an ASR system was trained using Sphinx [17]. 15 hours of audio from the DARPA Transtac Iraqi-English mediated interactions (San Diego corpus) [3] were used to train tied-state triphone HMM acoustic models with 1000 senones and 32 component Gaussian mixture models for every state. A trigram language model was trained on a large corpus of conversational English text [18] from the internet and other sources using the SRILM toolkit [19]. The WER of this system on a test set of 1000 utterances (distinct from the training set) was 29%.

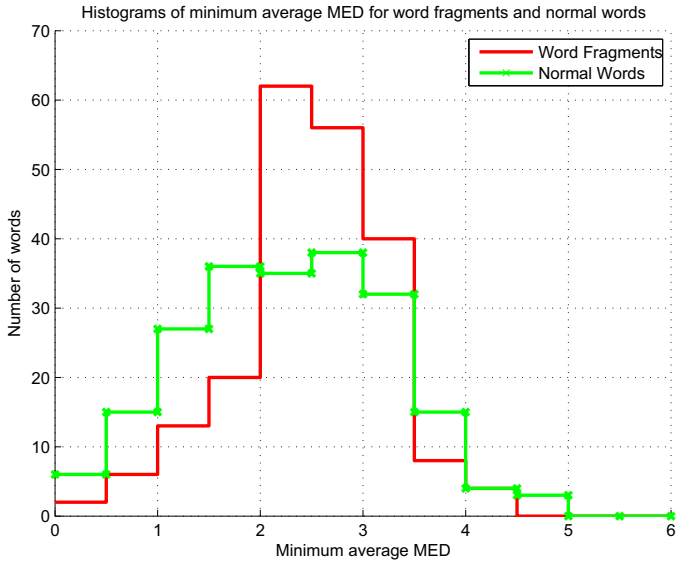

Fig. 3. Histograms of the minimum average MED across the ASR output lattice during word fragments and normal words. 211 samples each of word fragments and normal words were taken from the San Diego corpus [3].

TABLE I
LIST OF ACOUSTIC-PROSODIC AND LEXICAL FEATURES

No.	Acoustic-prosodic	Lexical
1	Change in avg. F0 from overall value	Min. no. unique hypotheses
2	Change in F0 across word boundary	Min. entropy
3	Log ratio of min. F0 before boundary to max. F0 after it	Min. hypotheses per word
4.	Change in avg. energy from overall value	Min. avg. MED
5	Change in energy across word boundary	Max. no. unique hypotheses
6	Jitter of the word	Max. entropy
7	Max. spectral tilt	Max. hypotheses per word
8	Min. spectral tilt	Max. avg. MED
9	Mean spectral tilt	Mean no. unique hypotheses
10	Change in energy slope across word boundary	Mean entropy
11	Max. OQ	Mean hypotheses per word
12	Min. OQ	Mean avg. MED
13	Mean OQ	

A. Classification Corpus

For conducting word fragment classification experiments, a set of files from the San Diego corpus containing 211 word fragments was extracted. It was noted that around 8.5% of the total words in this dataset were marked as fragments. Unlike the Switchboard corpus [10] used in [1], the dictionary for this corpus did not include the word fragments. This represents a more realistic situation in practice, since the word fragments are typically not known a-priori. Hence, for creating the oracle annotation for the classification corpus, the word fragments were removed from the transcripts of these files, and they were force aligned using the English ASR system. The resulting word segmentation files were manually edited by listening to the audio and marking the location of word fragments. It was found out that the WER of the ASR system on this small set of word fragment containing files was 40%. This high value was obtained since these files were not part of the text used to train the language model. The word fragments are OOV words, which leads to greater misrecognition.

Finally, word lattices were generated for this test set. The proposed lexical features were extracted from these lattices, as explained previously. The total number of lexical features was 12. For comparison with, and to complement our feature set, we also implemented the acoustic-prosodic features proposed in [1]. For the F0-based features, the pitch extractor tool in [20] was used. Praat [21] was used to extract the point process for computing jitter. A total of 13 acoustic-prosodic features were extracted. The final list of these features and the proposed lexical features is given in Table I.

B. Classification Experiments (without feature selection)

Since the number of normal words is very large as compared to the number of word fragments (2268 vs. 211), a classifier will get biased towards the majority class upon supervised training. We adopt the same solution proposed by Liu [1], and randomly downsample the set of normal words to 211. The LibSVM [22] Support Vector Machine (SVM) [23] classifier with a quadratic polynomial kernel was used for the experiments in the Weka 3 [24] data mining software. The random downsampling of normal words was repeated 10 times and the results were averaged across all the resulting test sets. A 10-fold cross validation was used for each of these 10 sets, with a train-test set split of $90 - 10\%$. Table II gives the average classification scores along with the standard deviation over these 10 randomly downsampled sets. The proposed lexical features outperform the acoustic-prosodic features by 5.20% (relative). The combination of the two sets of features leads to an improvement of 11.50% (relative) in the F measure over the acoustic-prosodic features alone. It must be noted that the classification scores of the three feature sets are significantly above the chance level of 0.50. Fig. 4 shows the Receiver Operating Characteristic (ROC) curves for these three sets of features using the same classification setup. The areas under the curve are 0.686, 0.728 and 0.773 for the acoustic-prosodic, lexical and combined feature set respectively. It can be observed that the combined feature set gives an appreciably higher true positive rate than the acoustic-prosodic features for all values of false positive rate.

By a simple concatenation of the two sets of features, we obtain a 25-dimensional vector. In the section IV.C, we discuss some feature selection experiments conducted on this combined feature set.

TABLE II
RESULTS OF CLASSIFICATION EXPERIMENT WITH ALL THE ACOUSTIC-PROSODIC, LEXICAL AND COMBINED FEATURES.

Score	**Acoustic-prosodic (all 13)**	**Lexical (all 12)**	**Combined (all 25)**
Precision	0.637 ± 0.018	0.671 ± 0.020	0.709 ± 0.020
Recall	0.636 ± 0.018	0.671 ± 0.020	0.708 ± 0.020
F measure	0.635 ± 0.019	0.668 ± 0.020	0.708 ± 0.020
% Improvement over acoustic-prosodic features	-	5.20%	11.50%

TABLE III
TOP 10 FEATURES SELECTED BY THE SVM ATTRIBUTE EVALUATION METHOD

Rank	**Acoustic-prosodic**	**Lexical**	**Combined**
1	Change in avg. energy from overall value	Mean entropy	Change in avg. energy from overall value
2	Jitter	Min. avg. MED between hypotheses	Mean entropy
3	Min. OQ	Mean no. of unique hypotheses	Max. spectral tilt
4	Max. spectral tilt	Min. entropy	Jitter
5	Change in energy across word boundary	Min. no. of unique hypotheses	Min. avg. MED between hypotheses
6	Change in F0 across word boundary	Max. no. of unique hypotheses	Log ratio of min. F0 before boundary to max. after it
7	Mean OQ	Max. avg. MED between hypotheses	Min. OQ
8	Change in energy slope across word boundary	Max. hypotheses per word	Min. entropy
9	Log ratio of min. F0 before boundary to max. after it	Mean avg. MED between hypotheses	Min. no. of unique hypotheses
10	Max. OQ	Mean hypotheses per word	Change in F0 across word boundary

C. Feature Selection and Classification

The small size of the database and the high dimensionality of the joint feature vector prompted us to conduct feature selection experiments. An additional advantage of this experiment is that it gives insights into the relative importance of each of the features for a classification task. The SVM Attribute Evaluation (`SVMAttributeEval`) method in Weka [24] was used with the `Ranker` search method for this task. This method evaluates the importance of a feature by using an SVM classifier [25]. Features are ranked according to the square of the weight assigned to them by the SVM. The top 10 features for the two sets separately and the combined set are given in Table III. One can note that out of the top 10 features in the combined feature set, only 4 belong to the lexical set. The higher performance of the top 10 lexical features as compared to the acoustic-prosodic features in spite of this can be attributed to a higher overall discrimination ability of individual lexical features.

Classification experiments were conducted by varying the number of top features selected from the combined set in a similar setup as section IV.B. Fig 5 shows the variation in the average F measure using the combined features after feature selection with the number of features selected. We can observe that for dimensionality greater than 12, the performance of the combined feature set is significantly better than the average value obtained using the entire set of acoustic-prosodic and lexical features (taken separately). In addition, the effect of the "curse of dimensionality" is not evident, since the performance of the combined feature set is nearly constant after a dimensionality of 12.

V. CONCLUSION

This paper presented ASR word lattice-based lexical features for the problem of word fragment detection in conversational speech. These features are based on the hypothesis that the confusion inherent in the word lattices generated by the ASR system can be exploited for detecting word fragments. Two sets of lexical features were proposed. The first one captures the word confusion between the various hypotheses at a given time instant in the lattice. The second set considers the confusion between the pronunciation of word hypotheses. Classification experiments with an SVM classifier show that these lexical features perform better than the previously proposed acoustic-prosodic features by around 5.20%. Furthermore, a combination of both these feature sets improves the word fragment detection accuracy by 11.50% relative to the acoustic-prosodic features. It must be noted that the ASR used to generate the word lattices had a WER of 40% on the set of sentences containing word fragments. In spite of this, the lexical features performed better than those based on

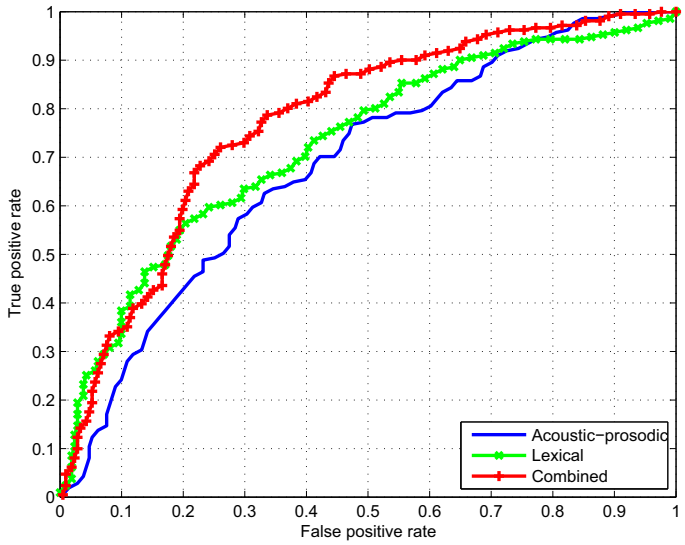

Fig. 4. ROC curves for classification experiments using acoustic-prosodic, lexical and combined feature sets (without dimensionality reduction). The areas under the curves are 0.686, 0.728 and 0.773 respectively.

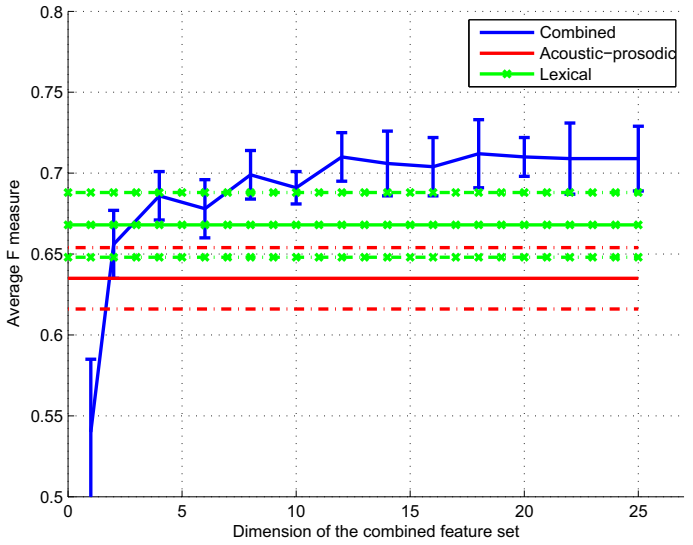

Fig. 5. Classification performance using a combined feature set of varying dimension. Also shown are the average F measures for the complete set of acoustic-prosodic and lexical features, with a ± 1 standard deviation margin.

acoustic-prosodic information. This highlights the robustness of the proposed features.

Our future work will investigate a richer corpus in term of acoustic variability. We will investigate the effect of regions of low reliability audio due to external factors. Finally, we would like to incorporate this work in a front end to a S2S system.

REFERENCES

[1] Y. Liu, "Word fragment detection using acoustic-prosodic features in conversational speech," in *Proc. HLT NAACL Student Research Workshop*, Edmonton, Canada, May 2003, pp. 37–42.

[2] C. Chu, Y. Sung, Y. Zhao, and D. Jurafsky, "Detection of word fragments in Mandarin telephone conversation," in *Proc. ICSLP*, Pittsburg, Pennsylvania, Sep. 2006.

[3] (2006) San Diego bilingual English-Arabic interpreter mediated interactions database. DARPA TRANSTAC Program.

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

[5] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006.

[6] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue," *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[7] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection," in *Proc. North American Chapter of ACL*, Boston, 2004, pp. 157–160.

[8] J. Bear, J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog," in *Proc. 30th Annual Meeting of the ACL*, 1992.

[9] C. Nakatani and J. Hirschberg, "A corpus-based study of of repair cues in spontaneous speech," *Journal of Acoustical Society of America*, pp. 1603–1616, 1994.

[10] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[11] S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Ganjavi, P. Georgiou, C. Hein, S. Kadambe, K. Knight, D. Marcu, H. Neely, N. Srinivasamurthy, D. Traum, and D. Wang, "Transonics: A speech to speech system for English-Persian interactions," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, Dec. 2003, pp. 670–675.

[12] E. Shriberg, "Phonetic consequences of speech disfluency," in *Proc. Intl. Conf. on Phonetic Sciences*, San Fransicso, California, 1999, pp. 619–622.

[13] A. E. Rosenberg, "The effect of glottal pulse shape on the quality of natural vowels," *Journal of Acoustical Society of America*, vol. 49, pp. 583–590, 1970.

[14] B. Blankenship, "The time course of breathiness and laryngealization in vowels," Ph.D. dissertation, Univ. of California, Los Angeles, 1997. [Online]. Available: http://www.linguistics.ucla.edu/faciliti/research/blankenship.pdf

[15] G. Fant, "The voice source in connected speech," *Speech Communication*, vol. 22, pp. 125–139, 1997.

[16] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, pp. 707–710, 1966.

[17] Sphinx 3. Carnegie Mellon University, Pittsburgh, Pennysylvania. [Online]. Available: http://cmusphinx.sourceforge.net/html/cmusphinx.php

[18] A. Sethy, P. Georgiou, and S. Narayanan, "Text data acquisition for domain-specific language models," in *Proc. EMNLP*, Sydney, Australia, 2006.

[19] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, Denver, Colorado, 2002, pp. 901–904.

[20] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, 1995, pp. 495–518.

[21] Praat: Doing phonetics by computer. P. Beorsma and D. Weenink. Version 5.1.05. [Online]. Available: http://www.praat.org

[22] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[23] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[24] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Fransicso: Morgan-Kaufmann, 2005.

[25] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.