IDENTIFICATION AND MODELING OF WORD FRAGMENTS IN SPONTANEOUS SPEECH

Yulia Tsvetkov, Zaid Sheikh, Florian Metze

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
{ytsvetko, zsheikh, fmetze}@cs.cmu.edu

ABSTRACT

This paper presents a novel approach to handling disfluencies, word fragments and self-interruption points in Cantonese conversational speech. We train a classifier that exploits lexical and acoustic information to automatically identify disfluencies during training of a speech recognition system on conversational speech, and then use this classifier to augment reference annotations used for acoustic model training. We experiment with approaches to modeling disfluencies in the pronunciation dictionary, and their effect on the polyphonic decision tree clustering. We achieve automatic detection of disfluencies with 88% accuracy, which leads to a reduction in character error rate of 1.9% absolute. While the high baseline error rates are due to the task we are currently working on, we demonstrate that this approach works well on the Switchboard corpus, for which the conversational nature of speech is also a major problem.

***Index Terms*---** speech recognition, conversational speech, word fragments identification, disfluency modeling, reference annotation

1. INTRODUCTION

Verbal disfluencies are phenomena that interrupt the flow of speech and do not add propositional content to an utterance. They include long pauses, repeated words or phrases, restarts, and revisions of content [1]. They constitute about 6% of word tokens in spontaneous speech, not including silent pauses [1, 2]. Liu examined conversations in Switchboard corpus [3, 4] and found that 17% of disfluencies are word fragments.

Because of their prevalence and irregularity, word fragments and other disfluencies degrade the performance of the speech recognizer: they can cause problems during training of the context decision tree, and the acoustic model (AM) itself. While hesitations, human noises and pauses can be modeled with dedicated models, word fragments are particularly problematic, because they are so similar to speech. They also add spurious content, and faulty linguistic structure to the language model training texts. Correct handling of disfluencies has been proven beneficial for various applications including language modeling [5], parsing of conversational speech [6], and speech-to-text translation [7].

We experiment with Cantonese and Turkish transcribed audio from the IARPA Babel Program [8] language collection releases babel101-v0.4c and babel105-v0.5. Cantonese speech transcripts contain 776,506 words, among them 10,107 word fragments. Although the percentage of word fragments is relatively small, there are 7.1% of utterances containing one or more fragments. In this work, we introduce lexical and acoustic features that help identify disfluencies in transcriptions of telephone conversations. Furthermore, unlike much existing work, which focused on disfluency removal from the ASR hypotheses or from the language model texts, our approach addresses incorporating disfluencies in the acoustic model training, and their representation in the pronunciation dictionaries. In addition, we combine our disfluency classifier and the best-scoring fragment modeling approach to adapt speech transcripts, and to discard noisy pronunciations from the acoustic model training.

After discussing related work in the next section, we describe our database in Section 3. In Section 4 we present the identification of word fragments, including a detailed discussion of the features and their implementation, as well as approaches to modeling disfluencies in pronunciation dictionary, and experiments with reference annotations update. Section 5 provides a thorough evaluation of the results. We conclude with suggestions for future research.

2. RELATED WORK

Liu et. al. [3, 9] train a decision tree classifier on multiple acoustic features to detect word fragments in the Switchboard corpus [4]. They extract prosodic features (duration, pitch tracks, and energy) and a set of voice quality measurements (jitter, spectral tilt, and open quotient) from forced alignments from speech to human transcriptions. They show that a prosody model alone performs at accuracy of 76.8% on the test data.

Chu et. al. [10] continue this line of work on the Mandarin corpus. They first replicate Liu's experiments [3] with acoustic features, but the results are quite different: they achieve 65.5% accuracy, and conclude that glottalization features are not as effective for Mandarin as for English. Rather than glottalization, they find that the most discriminative feature for Mandarin fragment detection is the identity of the neighboring word. The addition of 200 lexical binary features for the presence of one of the hundred most frequent words on the left or on the right of the word fragment candidate yielded an accuracy of 80%. In our experiments we rely on prosodic cues that have been proven beneficial in previous work. In addition, we introduce new lexical features, that capture additional properties of disfluencies, and significantly improve over the baseline of 80%.

To the best of our knowledge there has been no quantitative study on modeling of word fragments in the pronunciation dictionary, and incorporating them in the acoustic model training.

3. DATA

Experiments reported in this paper are performed as a part of the Babel project [8], whose goal is to develop speech recognition capability for keyword search in any language using limited amounts of transcribed speech. Our corpora include languages from a variety of language families, with diverse phonotactic, morphological, and syntactic characteristics. We use Cantonese babel101-v0.4c release as a training corpus, and Cantonese babel101-v0.4c development set and Turkish babel105-v0.5 corpora for evaluation. Audio data contain recordings from telephone conversations. The speakers are encouraged to talk about one of a pre-defined list of topics, including culture, sports, health and technology.

Cantonese corpus contains 158 hours of audio, with only about 71 hours of speech, produced by speakers from five different dialect groups identified by phonological, geographical and lexical variation. Turkish database contains 69 hours of audio (close to 40 hours of speech), also with 7 different speaker types. Such a diverse and low-resource setting makes it extremely challenging to build a good speech recognition system using state-of-the-art techniques.

In addition to demographic information, transcribed conversations may contain metadata on challenging acoustic conditions, hesitations, noises, mispronounced words and word fragments. Fragments are marked with a hyphen in the end, e.g.,你- ''you'' . We exploit this information in our experiments.

4. METHODOLOGY

4.1. Word Fragments Identification

To identify word fragments we train the SVM classifier on the training set of fragments and non-fragments extracted from the Cantonese reference transcripts. Positive examples are all word fragments that occur more than once in our training transcripts. Negative examples are complete words that have corresponding surface form in the positive examples (words from the set of positive examples, without the hyphen). Since non-fragment words are much more frequent, to create a balanced training set we downsample occurrences of each non-fragment to at most the number of corresponding word fragments occurrences. We extract 3724 positive and 3700 negative examples.

For acoustic features, audio frames of the candidates are obtained from Viterbi aligments of speech to the reference transcripts. Then, 1582 acoustic features [11] have been calculated on these audio frames using open-source extractor openSMILE [12]. Acoustic features correspond to various statistical functionals applied to PCM loudness, MFCC coefficients, log Mel Frequency Bands, line spectral pair frequencies, voicing probability and pitch-related features (F0, Jitter, Shimmer etc.) and ranked by information gain with the Weka toolkit [13].

Turn duration is the most salient feature. Features derived from voicing probability, 1st order delta coefficients of MFCC features, and smoothed LSP frequencies also rank in the top of the list. We experiment with several thresholds applied to the ranking, and with subsets of the extracted features containing only prosodic features related to pitch, energy and duration. Duration features include word duration and average phone duration in the word.

We also define the following lexical features aimed at capturing some of the unique properties of word fragments: frequency of the word in the training corpus, distance to the end of the turn, distance to the beginning of the turn, a binary feature whose value is 1 iff the word's location is before a noise model (to detect words cut by the phone line noises), a binary feature whose value is 1 iff the word is a prefix of the next word (to detect restarts).

4.2. Word Fragments Modeling

Suboptimal representation of word fragments can harm the alignment: although the number of word fragments in our data is only 1.3% percent, there are 7.1% of utterances containing one or more fragments. If a word fragment occurs in the reference transcriptions, but does not occur in the pronunciation dictionary, depending on the toolkit configuration, the Out-of-Vocabulary (OOV) word is either treated as a noise model, e.g., GARBAGE phone, or the whole utterance is discarded from training. There are standard techniques to handling OOV words, but word fragments have idiosyncratic acoustic properties, and therefore their dictionary representation should be different from corresponding complete words. However, we did not find any empirical study on modeling disfluencies in the dictionaries and incorporating them in AM training. We propose to incorporate word fragments in AM training by adding them to pronunciation dictionary, and investigate the effect of several different fragment representations. Each representation is a modification of the pronunciation of the corresponding complete word. Altering pronunciations affects the accumulated statistics of the phone and its context during training.

We detail below our experiments, and give examples of each approach.

GARBAGE phone instead of the last phone In partial words, we expect the word boundary to have irregular characteristics, hence, we replace the last phone of the word with GARBAGE phone, representing human noise model. For example, Cantonese word 上去 ‘‘go up’’ has the following pronunciation {{*s WB*} *9: N h* {*9y WB*}}, and corresponding word fragment 上去- is added as {{*s WB*} *9: N h* {*GARBAGE WB*}}.

FRAGMENT tag attached to the last phone FRAGMENT tag (FG) is a tag attached to phones (similar to a word boundary tag WB); here we attach the FG tag to the last phone. Different models will be trained for the same phone, depending on whether it is the last phone of a fragment, or of a complete word. It can be seen as a soft constraint, since if the pronunciation of a phone with tag is similar to the pronunciation of this phone without tag, they will be merged during polyphonic decision tree clustering. During context-dependent training the left context of the next word will be affected by the correct model. In this experiment, the pronunciation of the word 上去- is added as {{*s WB*} *9: N h* {*9y FG WB*}}.

GARBAGE phone instead of each phone In this configuration we replace all phones in the pronunciation of word fragment with the GARBAGE phone. Although we discard all fragments from training, this setup is different from simple treatment of the whole word as a GARBAGE phone, since we incorporate information on the length of the fragment, imposing the minimum number of garbage models in the Viterbi alignment of the phone transcriptions with the speech signal. 上去-, therefore, is modeled as {{*GARBAGE WB*} *GARBAGE GARBAGE GARBAGE* {*GARBAGE WB*}}

FRAGMENT tag attached to each phone In this experiment we attach FG tag to each phone, thereby training different models for the phones within word fragments. 上去- is pronounced as {{*s FG WB*} {*9: FG* } {*N FG* } {*h FG* } {*9y FG WB*}}.

In addition, we train two baseline systems: one uses word fragment pronunciation identical to the pronunciation of the complete word, in another we discard training utterances containing word fragments. We discuss experiments results in Section 5.

We discard information on word fragments during decoding: decoding dictionary contains only complete words and noise models. Hence, incomplete words and other verbal disfluencies are treated as human noise (GARBAGE) during recognition.

4.3. Adaptation of reference annotation

Annotation of word fragments in reference transcriptions is subjective and inconsistent. It depends on a listener who choses to mark repetition as a full word, e.g., 你 你 or as a fragment 你- 你, and on instructions to transcriber to mark fragments or not, to annotate them as a fragment word, as a complete word or as garbage. For example, in our Cantonese texts the bi-gram 你 你 ‘‘you you’’ occurs in all four possible configurations: 733 times in a form 你- 你, 874 times as 你 你, 167 times as 你- 你-, and 21 times as 你 你-. Such inconsistency in annotation not only affects acoustic models training, but also corrupts the linguistic structure of the language models training texts.

Therefore we run an experiment in which we update reference annotations according to predictions of the classifier presented in Section 4.1 and re-train the speech recognizer. Even if transcriptions contain non-fragment word, but our classifier marks it as fragment, we assume that it may contain disfluency (its too short, or it is a repetition of the next word, etc.) and we remove it from the training, thereby leaving only acoustically cleaner data for acoustic model training. We mark such word in the reference transcriptions as a fragment (appending a hyphen to a word), and add it as a word fragment to the pronunciation dictionary, with FRAGMENT tag attached to the last phone. Then, we re-train the system following the experimental setup detailed in the next Section.

5. RESULTS AND EVALUATION

To identify word fragments we train the SVM classifier (using LIBSVM [14]) with the radial basis function kernel. We use 3724 positive and 3700 negative examples extracted from Cantonese transcripts for training and evaluation: we perform 10-fold cross validation experiments, reporting accuracy of the word fragments classifier with several subsets of features described in Section 4.1. The results are depicted in Table 1. Top-ranked features correspond to 34 acoustic features ranked by their information gain (experimentation with several other thresholds yielded worse results). Prosodic features contain 41 features, these are all non-zero features corresponding to pitch and energy. The classification accuracy of prosodic features is very close to the result obtained by [10], suggesting that the word fragment acoustic properties are similar in Mandarin and Cantonese. However, our lexical features outperform frequency-based lexical features in [10] by 8.9%.

As a further demonstration of the utility of our approach, we classify Turkish feature vectors using Cantonese model. Following the methodology detailed in Section 4.1, we extract from Turkish reference transcripts 965 word fragments and 754 complete words, along with their feature vectors containing lexical, duration and prosodic features. We do not train the classifier, we instead use the Cantonese model weights to apply the classifier directly to the Turkish samples. The classification accuracy is 74.4%. While more careful evaluation is required in order to estimate the cross-lingual applicability of the classifier, we interpret this result as further proof of the robustness of our approach.

Features	Accuracy
Lexical	87.4%
Duration	57.8%
Prosodic	65.5%
Top-ranked	65.9%
Lexical+Duration+Prosodic	**88.8%**
Lexical+Duration+Top-ranked	88.7%
Lexical+Duration+Prosodic+Top-ranked	88.3%

Table 1. Classification accuracy of Cantonese training set in 10-fold cross-validation

To evaluate word fragments modeling in the pronunciation dictionary, we use the Cantonese corpus to train six acoustic models with different dictionaries, following four experimental setups discussed in Section 4.2, and two baseline setups. We train Maximum Likelihood (ML), context-dependent, fully-continuous systems with the JANUS Recognition Toolkit that features the IBIS single pass decoder [15]. In all systems we run 4 iterations of context-independent training, followed by one iteration of context-dependent training, there are 4K tied states, no speaker adaptation or discriminative training applied. Table 2 details the systems performance. Modeling fragments with FRAGMENT tag attached to the last phone is clearly helpful, improving the system accuracy by 4.4%.

Word fragments representation	CER
Baseline 1: discard utterances with word fragments	68.0%
Baseline 2: pronunciation of the fragment is identical to non-fragment	67.9%
GARBAGE phone instead of the last phone	68.0%
GARBAGE phone instead of each phone	68.2%
FRAGMENT tag attached to the last phone	**66.6%**
FRAGMENT tag attached to each phone	67.5%

Table 2. Speech recognition character error rate (CER) with different word fragment representations in Cantonese pronunciation dictionary

To validate our approach to modeling fragments, we measure the effect of alternative representations in the pronunciation of English word fragments in the Switchboard corpus. We use 300-hour Switchboard-I training set [16] to train six ML systems using the same training procedure as in the Cantonese experiments. We evaluate the performance of these systems on a 1-hour subset of the Eval02 data designed to have a similar error rate as the full Eval02 set [17]. We train each system with 7 iterations of context-independent training, followed by three iterations of context-dependent training with 10K tied states. Table 3 demonstrates that reduction in word error rates (WER) is consistent with our Cantonese experiments, and incorporating disfluencies in acoustic models training is beneficial for ASR systems.

Word fragments representation	WER
Baseline 1: discard utterances with word fragments	40.1%
Baseline 2: pronunciation of the fragment is identical to non-fragment	39.8%
GARBAGE phone instead of the last phone	39.4%
GARBAGE phone instead of each phone	39.6%
FRAGMENT tag attached to the last phone	**39.0%**
FRAGMENT tag attached to each phone	39.2%

Table 3. Speech recognition WER with different word fragment representations in English pronunciation dictionary

Finally, we evaluate the hypothesis that the disfluencies classifier can help augment reference transcripts, make them more consistent and remove noisy or irregular pronunciations from training. We replace only high-probability fragments in the Cantonese transcripts (words classified as fragments with confidence higher than 0.9). There are 24,678 classified fragments, comprising 3.2% of the references. The obtained character error rate is 66.1%, which is lower than the performance of our best system with the same dictionary configuration by 0.5% absolute. The adaptation of the transcripts is highly significant, improving the accuracy of the baseline by 6% (1.5% improvement over the best Cantonese system).

6. CONCLUSIONS AND FUTURE WORK

In this work we introduce lexical and acoustic features that improve automatic extraction of word fragments over the existing baselines. We also show how to handle disfluencies in pronunciation dictionary, to augment the HMM models of word fragment constituent phones and their context. In addition, we describe a novel method that combines our disfluency classifier and word fragment modeling approach to adapt speech transcripts, and to discard noisy pronunciations from the acoustic model training. In the future we are planning to continue experiments with adaptation of reference transcriptions and cross-lingual analysis of the disfluencies classifier.

7. ACKNOWLEDGEMENTS

We are grateful to Yajie Miao and Yuchen Zhang for their help.

Partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

8. REFERENCES

[1] J. E. Fox Tree, ``The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech,'' in *Journal of Memory and Language*, 1995, vol. 34, pp. 709--738.

[2] S. V. Kasl and G. F. Mahl, ``The relationship of disturbances and hesitations in spontaneous speech to anxiety,'' in *Journal of Personality and Social Psychology*, 1965, vol. 1(5), pp. 425--433.

[3] Y. Liu, ``Word fragment identification using acoustic-prosodic features in conversational speech,'' in *HLT-NAACL student research workshop*, 2003, pp. 37--42.

[4] J. Godfrey, E. Holliman, and J. McDaniel, ``SWITCHBOARD: Telephone speech corpus for research and development,'' in *Proc. ICASSP*, 1992, vol. 1, pp. 517--520.

[5] A. Stolcke and E. Shriberg, ``Statistical language modeling for speech disfluencies,'' in *Proc. ICASSP*, 1996.

[6] M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart, et al., ``Parsing speech and structural event detection,'' in *JHU Summer Workshop Final Report*, 2005.

[7] S. Rao, I. Lane, and T. Schultz, ``Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts,'' in *Machine Translation Summit XI*, 2007.

[8] ``IARPA - Solicitations - Office of Incisive Analysis, Babel Program,'' `http://www.iarpa.gov/solicitations_babel.html`, Retrieved 2012-04-23.

[9] Y. Liu, E. Shriberg, and A. Stolcke, ``Automatic disfluency identification in conversational speech using multiple knowledge sources,'' in *Proceedings of Eurospeech*, 2003.

[10] C. Chu, Y. Sung, Z. Yuan, and D. Jurafsky, ``Detection of word fragments in Mandarin telephone conversation,'' in *International Conference on Spoken Language Processing*, 2006.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, ``The INTERSPEECH 2010 paralinguistic challenge,'' in *Proc. INTERSPEECH*, 2010, pp. 2794--2797.

[12] F. Eyben, M. Wöllmer, and B. Schuller, ``openSMILE: the munich versatile and fast open-source audio feature extractor,'' in *Proceedings of the international conference on Multimedia*. 2010, pp. 1459--1462, ACM.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, ``The WEKA data mining software: an update,'' *SIGKDD Explorations*, vol. 11, no. 1, pp. 10--18, 2009.

[14] C. Chang and C. Lin, ``LIBSVM: A library for support vector machines,'' *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1--27:27, 2011.

[15] H. Soltau, F. Metze, C. Fügen, and A. Waibel, ``A one-pass decoder based on polymorphic linguistic context assignment,'' in *Proc. ASRU*, 2001.

[16] J.J. Godfrey and E. Holliman, ``Switchboard-1 release 2,'' *Linguistic Data Consortium, Philadelphia*, 1997.

[17] H. Soltau, H. Yu, F Metze, C. Fügen, Q. Jin, and S. Jou, ``The 2003 ISL rich transcription system for conversational telephony speech,'' in *in ICASSP*, 2004.