A Study On Natural Expressive Speech: Automatic Memorable Spoken Quote Detection

Fajri Koto†‡, Sakriani Sakti†, Graham Neubig†, Tomoki Toda†, Mirna Adriani‡, and Satoshi Nakamura†

Abstract This paper presents a study on natural expressive speech during public talks. Specifically, we focus on how people convey important messages that may be retained in the audience's consciousness. Our study aims to answer several questions. Why are some public speeches memorable and inspirational for the audience, while others are not? Why are some memorable/inspirational spoken quotes more popular than others? Being able to evaluate why certain spoken words are memorable/inspirational is not a trivial matter, and most studies on memorable quote detection are only limited to textual data. In this study, we use both linguistic and acoustic features of public speeches in TED talks. The results reveal that based on those linguistic and acoustic features, we are able to distinguish memorable spoken quotes and non-memorable spoken quotes with 70.4% accuracy. Furthermore, we also analyze the important factors that affect the memorableness and popularity of spoken quotes.

1 Introduction

Research related to spoken dialog systems has progressed from the traditional task-based frameworks to more sophisticated social agents [1] that can engage the user and expressively convey the intended message. Consequently, understanding the ways humans express themselves and engage their listeners is becoming more a more important factor in designing these sorts of systems. Here, we focus on studying natural expressiveness and its effects during public speeches.

Through history, the best speeches of all time normally feature memorable quotes that genuinely inspire the audience. For instance, the most famous quote of John F. Kennedy: *"Ask not what your country can do, ask what you can do for your country"*, has inspired many generations since he gave this speech in January 1961[1]. More recent examples of inspirational public speech can be found on TED[2]. TED features talks of 5-25 minutes by skilled speakers on subjects including Technology, Entertainment, Design. Many famous people have given speeches on TED and inspired people by their memorable words. Recently, TED has started "TED Quotes,"

F. Koto · S. Sakti · G. Neubig · T. Toda · S. Nakamura
†Nara Institute of Science and Technology (Japan) e-mail: {ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

F. Koto · M. Adriani
‡University of Indonesia (Indonesia) e-mail: fajri91@ui.ac.id, mirna@cs.ui.ac.id

[1] http://www.ushistory.org/

[2] http://www.ted.com/

which collects memorable quotes from TED talks, annotates them manually, groups them by category, and provides an easy way for people to share their favorite quotes. The most popular quotes can have more than a thousand shares.

While some public speeches may have inspired many individuals, they raise deeper questions. Why are some spoken words be memorable and inspirational, while some others are not? Why are some memorable quotes more popular than others? Answering these questions will be more challenging than just determining whether particular keywords appear in a given segment of speech as in spoken term detection research [2, 3]. Memorable quote detection involves the evaluation of what is being said by the speaker and how the audience reacts, even with or without particular keywords. The challenge lies in detecting generic pearls of wisdom expressed with unusual combinations of words.

We argue that there may be specific patterns or combination of words, as well as specific intonation or accent patterns which distinguish memorable spoken quotes from other spoken utterances. In this study, we attempt to answer these questions by developing a method for automatic detection of memorable spoken quotes and analyzing their popularity.

2 Memorable Spoken Quotes Detection

Research related to memorable quote detection is still very limited. Bandersky et al. proposed an automatic detection of memorable quotes from books using linguistic features [4]. Research by Kolak et al. also proposed an approach for automatically mining quotes from extremely large text corpora [5]. Similar work by Liang et al. automatically extracts quotations and allows for efficient retrieval of the semantically annotated quotes from news stories [6]. Another study by Danescu-Niculescu-Mizil et al. attempted to investigate the effect of phrasing on a quote's memorability from movie scripts [7]. While most techniques developed so far for memorable quote detection have focused primarily on the processing of text, we are interested in discovering memorable spoken quotes from natural speech.

2.1 Corpus Construction

To enable the system to learn to distinguish between memorable and non-memorable spoken quotes, we compiled a corpus from the TED website. The collected memorable quotes resulted in a total of 2118 speech transcription segment files. To construct a corpus for comparison, we also randomly selected a total of 2118 speech transcription segment files from the rest of the data and labeled them as non-memorable quotes.

Within TED, there is a "share" function that allows users to share their favorite quotes with others, and we utilize the number of shares as a measure of popularity. Here, we only focused on extreme cases and constructed a corpus with memorable quotes that have zero shares (labeled as non-popular quotes), and memorable quotes that have more than 50 shares (labeled as popular quotes). Here, all newly published quotes still have zero shares, and thus we exclude them from the data. In total, the corpus consists of 262 non-popular quotes and 179 popular quotes.

Further details of data construction can be found in our previous work [8].

2.2 Features of Spoken Quotes

Bandersky et al. defined three kinds of linguistic features useful for memorable quote detection: lexical, punctuation, and part-of-speech (POS) [4]. Following these linguistic features, we utilize lexical features (*#capital*, *#quantifier*, *#stops*, *beginStop*, *hasDialog*, *#abstract*) and part-of-speech (*#POS*, *hasComp*, *hasSuper*, *hasPP*, *#IGSeq[i]*) features. As we focus on spoken utterances of memorable quotes, punctuation features are excluded. In addition, we included *hasSynonym* and *hasAntonym* features in our experiment. A detailed descriptions of these features are shown in Table 1. For *#quantifier*, *#stop*, and *#abstract* features, we use 17 quantifiers[3], 174 stop words[4], and 176 abstract words[5], respectively.

The *#IGSeq[i]* feature is used to analyze the pattern of POS sequences. Here, we generate feature of tri-POS sequences from the data, resulting in 5724 generated POS sequences. We then computed the information gain of all POS sequences in all memorable and non-memorable quotes based on Equation (1) and Equation (2),

$$IG(X,Y) = H(X) - H(X|Y) \tag{1}$$

$$H(X) = -p(x)log_2 p(x). \tag{2}$$

Feature *#IGSeq[i]* expresses the number of times the i-th POS sequence is contained in quote s, where X indicates the presence or absence of the POS sequence in current quote, and Y indicates the type of quote (memorable or non-memorable). In this study, based on the information gain of all POS sequences, we selected only the top-250 of POS sequences as linguistic features.

While previous work has focused on lexical features, in this study we also include acoustic features. Specifically, we use the INTERSPEECH 2010 paralinguistic challenge configuration (IS10 Paraling features) [12]. It consists of 1582 features, which are obtained in three steps: (1) 38 low-level descriptors are extracted and smoothed by simple moving average low-pass filtering; (2) their first order regression coefficients are added; (3) 21 functionals are applied. However, 16 zero-information features (e.g. minimum F0, which is always zero) are discarded. Finally, two single features for F0: number of onsets and turn duration are added. More details of each feature can be found in [12] and [14].

2.3 Classifier

Based on this corpus, we develop a method for automatic detection of memorable spoken quotes. Specifically, we use both linguistic and acoustic features to distinguish between memorable quotes and non-memorable quotes of public speeches in TED talks. We investigated three classifiers: Neural Networks (NN)[9], Naive Bayes

[3] http://www.tesol-direct.com/guide-to-english-grammar/quantifiers

[4] http://www.ranks.nl/resources/stopwords.html

[5] http://www.englishbanana.com

Table 1 Linguistic feature sets for a particular quote *s*.

Feature	Description
	Lexical
#capital	Number of capitalized words in *s*
#quantifier	Number of universal quantifiers in *s*
#stops	Number of common stopwords in *s*
beginStop	True if *s* begins with a stopword, False otherwise
hasDialog	True if *s* contains at least one of say, says, said
#abstract	Number of abstract concepts (e.g., adventure, charity, stupidity) in *s*
	Part of Speech
#POS	POS = noun, verb, adjective, adverb, pronoun
hasComp	True if *s* contains a comparative adjective or adverb, False otherwise
hasSuper	True if *s* contains a superlative adjective or adverb, False otherwise
hasPP	True if *s* contains a verb in past participle, False otherwise
hasSynonym	True if *s* contains two words that are synonymous, False otherwise
hasAntonym	True if *s* contains two words are antonyms of each other, False otherwise
#IGSeq[i]	Count of the POS sequence with *i*-th highest IG(X,Y) (Eq.1) in *s*

(NB)[11], and Support Vector Machines (SVM) [10]. We also performed feature selection with forward algorithm approach to estimate the best feature set.

3 Experimental Set-Up and Evaluation

3.1 Set-up

Linguistic features were extracted using NLTK [13], while acoustic features were extracted using openSMILE toolkit[6] [14]. There are a total of 264 linguistic features and 1582 acoustic features. Here, we perform 5-fold cross validation with 80% of the corpus as training set, with the remainder of the corpus as the test set. Training of the prediction models was performed with Rapidminer[7][15].

3.2 Memorable Quote Detection

First, we conducted memorable quote detection for all features and classifiers (NN, NB, and SVM). Table 2 shows the performance of all classifiers after feature selection. As a comparison, we also include the performance of the classifier using the top-10 features of memorable quotes detection proposed by Bandersky which are obtained by SVM weighting (denoted as "Baseline"). The results reveal that our proposed features give better accuracy than the baseline, and the best results were achieved by the use of acoustic features.

Next, we combine selected features from all classifiers into one union set of selected features. As there are some overlap of features, we finally have 12 linguistic features and 9 acoustic features in total. The result shows the accuracy of memorable quote detection based on an SVM classifier, using: (1) 12 selected linguistic features

[6] http://opensmile.sourceforge.net/

[7] http://www.rapidminer.com

Table 2 Accuracy of memorable quote detection with 5-fold cross validation for baseline and proposed features (the chance rate is 50.0%).

	Baseline	Proposed	
Classifier	Linguistic	Linguistic	Acoustic
Neural Network	63.98%	64.87%	**67.71**%
Naive Bayes	62.91%	65.04%	**68.18**%
Support Vector Machine	64.80%	66.71%	**68.08**%

Table 3 POS-tag sequences selected for memorableness analysis (MQ = Memorable Quotes and NM = Non-Memorable Quotes)

Sequence	Example	#MQ	#NM
CC-PRP-VBD	but i thought, and i introduced	43	124
NN-VBZ-DT	belief is the, education is a	155	45
JJ-NN-NN	national automobile slum, quiet screaming desperation	236	183
PRP-VBZ-IN	it is as, it is like	95	39
NN-VBZ-RB	innovation is not, privacy is not	165	50

only with 66.45% accuracy, (2) 9 selected acoustic features only with 68.06%, and (3) combination of the selected linguistic and acoustic features with the highest, 70.4% accuracy. The results reveal that the classifier with all features performs better than the classifier with linguistic or acoustic features only.

3.3 Memorableness and Popularity Analysis

We further analyze the features selected by the feature selection procedure. For acoustic features, the selected features are mainly F0, logMelFreqBand, and MFCC. By performing SVM weighting on these selected features, we found out that F0 had the highest weight. It indicates that the prosody of the utterance is a significant feature that distinguishes between memorable quotes and non-memorable quotes.

For linguistic features, the selected features include beginStop, #noun, #adjective and some POS-tag sequences. The details of those POS-tag sequences including examples of word sequences are given in Table 3. CC-PRP-VBD is actually an amalgamation of two single sentences, a compound sentence. Based on Table 3, the sentences containing CC-PRP-VBD sequences tend to be non-memorable quotes. This indicates that memorable quotes seldom use conjunctions or they usually consist of single sentences. On the other hand, sentences with POS sequences of NN-VBZ-DT, JJ-NN-NN, PRP-VBZ-IN and NN-VBZ-RB tend to be memorable quotes. These POS sequences are mainly used for definition, elaboration and explanation types of sentences. Based on this data, we may argue that memorable quotes tend to contain general statements about the world from the perspective of the speaker.

For the popularity analysis, the experiment was conducted utilizing only linguistic features, as people share their favorite quotes based only on text. Our highest classification result was achieved by Naive Bayes with **69.40%** accuracy. The accuracy of Neural Network and SVM are 68.48% and 62.13% respectively.

4 Conclusion

In this study, we discussed the possibilities of automatically detecting the memorable spoken quotes in real public speeches based on linguistic and acoustic features. The results reveal that a classifier with both linguistic and acoustic features performs better than a classifier with linguistic or acoustic features only. By the use of this features combination, we can distinguish between memorable quotes and non-memorable quotes with 70.4% accuracy. Based on the analysis of the selected features, the results reveal that most memorable quotes have definition, elaboration and explanation type sentences, and the prosody of utterances is a significant acoustic feature that distinguishes between memorable quotes and non-memorable quotes.

Acknowledgements Part of this work was supported by JSPS KAKENHI Grant Number 26870371.

References

1. Dautenhahn K (2007) Socially intelligent robots: dimensions of humanrobot interaction. Phil. Trans. R. Soc. B, vol.362:679-704,
2. Miller D R, Kleber M, Kao C L, Kimball O, Colthurst T, Lowe S A, Gish H (2007) Rapid and accurate spoken term detection. In: Proc. INTERSPEECH:314–317
3. Vergyri D, Shafran I, Stolcke A, Gadde V R R, Akbacak M, Roark B, Wang W (2006) The SRI/OGI 2006 spoken term detection system. In: Proc. INTERSPEECH:2393–2396
4. Bandersky M, Smith, D A (2012) A Dictionary of Wisdom and Wit: Learning to Extract Quotable Phrase. In: Proc. NAACHL-HLT, Montréal, Canada:69–77
5. Kolak O, Schilit B N (2008) Generating links by mining quotations. In: Proc. 9th ACM conference on Hypertext and hypermedia:117–126
6. Liang J, Dhillon N, Koperski K (2010) A large-scale system for annotating and querying quotations in news feeds. In: Proc. 3rd International Semantic Search Workshop:7
7. Danescu-Niculescu-Mizil C, Cheng J, Kleinberg J, Lee L (2012) You had me at hello: How phrasing affects memorability. In: Proc. ACL, Jeju Island, Korea:892–901
8. Koto F, Sakti S, Neubig G, Toda T, Adriani M, Nakamura S. (2014) Memorable Spoken Quote Corpora of TED Public Speaking. In: Proc. the 17th Oriental COCOSDA, Phuket, Thailand:140–143
9. Ful L (1994) Neural Network in Computer Intelligence. McGraw-Hill International Edition, MIT-Press.
10. Lewis D D (1998) Naive Bayes at forty: The independence assumption in information retrieval. In: Proc. ECML-98, Berlin Heidelberg:4–15
11. Cristianini N, Taylor J S (2000) An Introduction to Support Vector Machines and Other kernel-based learning methods. Cambridge University Press
12. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Muller C A, Narayanan S S (2010) The INTERSPEECH 2010 Paralinguistic Challenge. In Proc. INTERSPEECH, Makuhari, Japan: 2794-2797
13. Bird S(2006) NLTK: The Natural Language Toolkit. In: Proc. COLING/ACL on Interactive Presentation Sessions, Sydney, Australia:69–72
14. Eyben F, Woeller M, Schuller B (2010) openSMILE – The Munich versatile and fast open-source audio feature extractor. In: Proc. Multimedia (MM):1459–1462
15. Akthar F, Hahne C (2012) RapidMiner 5 Operator Reference. Rapid-I GmbH.