Towards Style Transformation from Written-Style to Audio-Style

Amjad Abu-Jbara[*]
EECS Department
University of Michigan
Ann Arbor, MI, USA
amjbara@umich.edu

Barbara Rosario
Intel Labs
Santa Clara, CA, USA
barbara.rosario@intel.com

Kent Lyons
Intel Labs
Santa Clara, CA, USA
kent.lyons@intel.com

Abstract

In this paper, we address the problem of optimizing the style of textual content to make it more suitable to being listened to by a user as opposed to being read. We study the differences between the written style and the audio style by consulting the linguistics and journalism literatures. Guided by this study, we suggest a number of linguistic features to distinguish between the two styles. We show the correctness of our features and the impact of style transformation on the user experience through statistical analysis, a style classification task, and a user study.

1 Introduction

We live in a world with an ever increasing amount and variety of information. A great deal of that content is in a textual format. Mobile technologies have increased our expectations as to when, where, and how we can access such content. As such, it is not uncommon to want to gain access to this information when a visual display is not convenient or available (while driving or walking for example). One way of addressing this issue is to use audio displays and, in particular, have users listen to content read to them by a speech synthesizer instead of reading it themselves on a display.

While listening to speech opens many opportunities, it also has issues which must be considered when using it as a replacement for reading. One important consideration is that the text that was originally written to be read might not be suitable to be listened to. Journalists, for example, write differently for audio (i.e. radio news broadcast) compared to writing content meant to be read (i.e. newspaper articles) (Fang, 1991).

One key reason for the difference is that *understanding* is more important than grammar to a radio news writer. Furthermore, audio has different perceptual and information qualities compared to reading. For example, the use of the negations *not* and *no* should be limited since it is easy for listeners to miss that single utterance. Listener cannot relisten to a word; and, missing it has a huge impact on meaning.

In this paper, we address the problem of changing the writing-style of text to make it suitable to being *listened to* instead of being read.

We start by researching the writing-style differences across text and audio in the linguistics and journalism literatures. Based on this study, we suggest a number of linguistic features that set the two styles apart. We validate these features statistically by analyzing their distributions in a corpus of parallel text- and audio-style documents; and experimentally through a style classification task. Moreover, we evaluate the impact of style transformation on the user experience by conducting a user study.

The rest of this paper is organized as follows. In the next section, we examine the related work. In Section 3, we summarize the main style differences as they appear in the journalism and linguistics literatures. In Section 4, we describe the data that we collected and used in this work. The features that we propose and their validation are discussed in Section 5. In Section 6, we describe the user study and discuss the results. We conclude in Section 7.

2 Related Work

There has been a considerable amount of research on the language variations for different registers and

[*] Work conducted while interning at Intel Labs

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 248–253,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

genres in the linguistics community, including research that focused on the variations between written and spoken language (Biber, 1988; Halliday, 1985; Esser, 1993; Whittaker et al., 1998; Esser, 2000). For example, Biber (1988) provides an exhaustive study of such variations. He uses computational techniques to analyze the linguistic characteristics of twenty-three spoken and written genres, enabling identification of the basic, underlying dimensions of variation in English.

Halliday (1985) performs a comparative study of spoken and written language, contrasting the prosodic features and grammatical intricacy of speech with the high lexical density and grammatical metaphor or writing. Esser (2000) proposes a general framework for the different presentation structures of medium-dependent linguistic units.

Most of these studies focus on the variations between the written and the *spontaneous* spoken language. Our focus is on the *written* language for audio, i.e. on a style that we hypothesize being somewhere between the formally written and spontaneous speech styles. Fang (1991) provides a pragmatic analysis and a side-by-side comparisons of the "writing style differences in newspaper, radio, and television news" as part of the instructions for journalist students learning to write for the three different mediums.

Paraphrase generation (Barzilay and McKeown, 2001; Shinyama et al., 2002; Quirk et al., 2004; Power and Scot, 2005; Zhao et al., 2009; Madnani and Dorr, 2010) is related to our work, but usually the focus has been on the semantics, with the goal of generating relevant content, and on the syntax to generate well formed text. In this work the goal is to optimize the style, and generation is one approach to that end (we plan addressing it for future work)

Authorship attribution (Mosteller and Wallace, 1964; Stamatatos et al., 2000; Argamon et al., 2003; Argamon et al., 2007; Schler and Argamon, 2009) is also related to our work since arguably different authors write in different styles. For example, Argamon et al. (2003) explored differences between male and female writing in a large subset of the British National Corpus covering a range of genres. Argamon el al. (2007) addressed the problem of classifying texts by authors, author personality, gender of literary characters, sentiment (positive/negative feeling), and scientific rhetorical styles. They used lexical features based on taxonomies of various semantic functions of different lexical items (words or phrases). These studies focused on the correlation between style of the text and the personal characteristics of its author. In our work, we focus on the change in writing style according to the change of the medium.

3 Writing Style Differences Across Text and Audio

In this section, we summarize the literature on writing style differences across text and audio. Style differences are not due to happenstance. Writing styles for different media have evolved due to the unique nature of each medium and to the manner in which its audience consumes it. For example, in audio, the information must be consumed sequentially and the listener does not have the option to skip the information that she finds less interesting.

Also, the listener, unlike the reader, cannot stop to review the meaning of a word or a sentence. The eye skip around in text but there is not that option with listening. Moreover, unlike attentive readers of text, audio listeners may be engaged in some task (e.g. driving, working, etc.) other than absorbing the information they listen to, and therefore are paying less attention.

All these differences of the audio medium affect the length of sentences, the choice of words, the structure of phrases of attribution, the use of pronouns, etc.

Some general guidelines of audio style (Biber, 1988; Fang, 1991) include 1) the choice of simple words and short, declarative sentences with active voice preferred. 2) Attribution precedes statements as it does in normal conversations. 3) The subject should be as close to the predicate as feasible. 4) Pronouns should be used with a lot of wariness. It is better to repeat a name, so that the listener will not have to pause or replay to recall. 5) Direct quotations are uncommon and the person being quoted is identified before the quotation. 6) Dependent clauses should be avoided, especially at the start of a sentence. It is usually better to make a separate sentence of a dependent clause. 7) Numbers should be approximated so that they can be under-

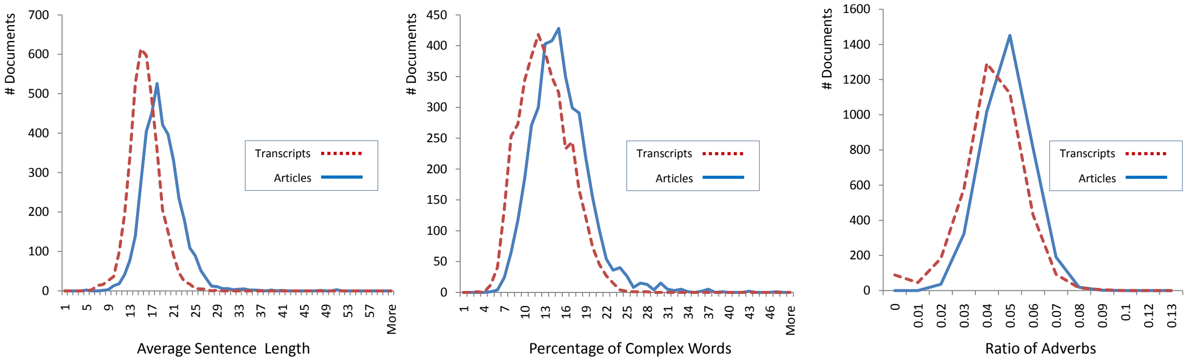

Figure 1: The distributions of three features for both articles and transcripts

stood. For example, the sum of $52,392 could be stated as *more than fifty thousand dollars*. 8) Adjectives and adverbs should be used only when necessary for the meaning.

4 Data

In order to determine the differences between the text and audio styles, we needed textual data that ideally covered the same semantic content but was produced for the two different media. National Public Radio (NPR) has exactly this type of data. Through their APIs we obtained the same semantic content in the two different styles: written text style (articles, henceforth) and in audio style (transcripts, henceforth). The NPR Story API output contains links to the Transcript API when a transcript is available. With the Transcript API, we were able to get full transcripts of stories heard on air[1]. To the best of our knowledge, this is the first use of this collection for NLP research.

We collected 3855 news articles and their corresponding transcripts. The data cover a varied set of topics from four months of broadcast (from March 6 to June 3, 2010). Table 2 shows an example of such article-transcript pairs.

5 Features

Based on the study of style differences outlined in section 3, we propose a number of document-level, linguistic features that we hypothesized distinguish the two writing styles. We extracted these features for each article and transcript. The analysis of these features (will be discussed later in the section) showed that they are of different importance to style identification. Table 1 shows a list of the top features and their descriptions.

5.1 Statistical Analysis

The goal of this analysis is to show that the values of the features that we extracted are really different across the two styles and that the difference is significant. We compute the distribution of the values of each feature in articles and its distribution in transcripts. For example, Figure 1 shows the distributions of 3 features for both articles and transcripts. The figure clearly shows how the distributions are different. A two-tailed paired Student's T-test (with alpha set to 0.05) reveals statistically significant difference for all of the features ($p < 0.0001$).

This analysis corroborated our linguistic hypotheses, such as the average sentence length is longer for articles than for transcripts, complex words (more than 3 syllables) are more common in articles, articles contain more adverbs, etc.

5.2 Classification

To further verify that our features really distinguish between the two writing styles, we conducted a classification experiment. We used the features described in Table 1 (excluding the *Direct Quotation* feature) and the dataset described in section 4 to train a classifier. We used Libsvm (Chang and Lin, 2001) with a linear kernel as our classifier. We performed 10-fold cross validation on the entire dataset.

[1] http://www.npr.org/api/index

Feature	Description	Rank
Direct quotations	We use a pattern matching rule to find all the instances of direct speech (e.g. ”I love English”, says Peter).	1
Average sentence length	The length of a sentence is the number of words it contains.	2
Ratio of complex words	A complex word consists of three or more syllables (Gunning, 1952). Complex words are more difficult to pronounce and harder to understand when being listened to than simpler words.	3
Ratio of pronouns	We count the different types of pronouns; first person pronouns, second person pronouns, third person pronoun, demonstrative pronouns (this, these, those), and the pronoun it.	4
Average distance between each verb and its subject	We associate each verb with its subject by parsing the sentence using a dependency parser and finding $nsubj$ link. The distance is the word count between the verb and its subject.	5
Ratio of adjectives	We count attributive adjectives (e.g. the big house) and predictive adjectives (e.g. the house is big) separately.	6
Dependent clauses	We identify dependent clauses by parsing the sentence and finding a $SBAR$ node in the parse tree.	7
Average noun phrase modification degree	The average number of modifiers for all the noun phrases in the document.	8
Average number of syllables	The total number of syllables in the document divided by the number of words. To get an accurate count of syllables in a word, we look up the word in a dictionary. All the numbers are converted to words (e.g. 25 becomes *twenty five*). We also change all the contractions to their normal form (e.g. *I'll* becomes *I will*).	9
Ratio of passive sentences	We find passive sentences using a pattern match rule against the part-of-speech tags of the sentence. We compute the ratios of agentless passive sentences and by-passive sentences separately.	10
Ratio of adverbs	In addition to counting all the adverbs, we also count special types of adverbs separately including: amplifiers (e.g. absolutely, completely, enormously, etc), downtoners (e.g. almost, barely, hardly, etc), place adverbials (e.g. abroad, above, across, etc), and time adverbials (e.g. afterwards, eventually, initially, etc). The list of special adverbs and their types is taken from Quirk et. al (1985).	11
Size of vocabulary	The number of unique words in a document divided by the total number of words.	12
Ratio of verb tenses	We count the three main types of verbs, present, past, and perfect aspect.	13
Ratio of approximated numbers	We count the instance of approximated numbers in text. In particular, we count the pattern *more than/less than/about/almost ¡integer number¿*.	14

Table 1: Style Features

Written article	Transcript
The mammoth oil spill in the Gulf of Mexico, sparked by the explosion and sinking of a deep-water oil rig, now surrounds the Mississippi River Delta, all but shutting down fisheries. But the oil industry still has a lot of friends on the delta. As Louisianans fight the crude invading their coast, many also want to repel efforts to limit offshore drilling. “We need the oil industry, and down here, there are only two industries – fishing and oil,” says charter boat captain Devlin Roussel. Like most charter captains on the delta, Roussel has just been sitting on the dock lately. But if he did have paying customers to take out fishing, he'd most likely take them to an oil rig. [..]	It's MORNING EDITION from NPR News. I'm Steve Inskeep. And I'm Renee Montagne. President Obama's administration is promising action on that catastrophic oil spill. The president's environmental adviser says the BP oil leak will be plugged. More on that in a moment. President Obama yesterday said the nation is too dependent on fossil fuels. But you dont realize just how dependent until you travel to the Mississippi River Delta. The fishing industry there is all but shut down. Yet some residents do not want to stop or slow offshore drilling despite the disaster. NPR's Frank Morris visited Buras, Louisiana [..]

Table 2: An example of an article–transcript pair.

Our classifier achieved 87.4% accuracy which is high enough to feel confident about the features.

We excluded the *Direct Quotation* feature from this experiment because it is a very distinguishing feature for articles. The vast majority of the articles in our dataset contained direct quotations and none of the transcripts did. When this feature is included, the accuracy rises to 97%.

To better understand which features are more important indicators of the style, we use Guyon et al.'s (2002) method for feature selection using SVM to rank the features based on their importance. The ranks are shown in the last column in Table 1.

6 User Study

Up to this point, we know that there are differences in style between articles and transcripts, and we formalized these differences in the form of linguistic features that are easy to extract using computational techniques. However, we still do not know the impact of changing the style on the user experience. To address this issue, we did manual transformation of style for 50 article paragraphs. The transformation was done in light of the features described in the previous section. For example, if a sentence is longer than 25 words, we simplify it; and, if it is in passive voice we change it to active voice whenever possible, etc. We used a speech synthesizer to convert the original paragraphs and their transformed versions into audio clips. We used these audio clips to conduct a user study.

We gave human participants the audio clips to listen to and transcribe. Each audio clip was divided into segments 15 seconds long. Each segment can be played only once and pauses automatically when it is finished to allow the user to transcribe the segment. The user was not allowed to replay any segment of the clip. Our hypothesis for this study is that audio clips of the transformed paragraphs (audio style) are easier to comprehend, and hence, easier to transcribe than the original paragraphs (text style). We use the edit distance between the transcripts and the text of each audio clip to measure the transcription accuracy. We assume that the transcription accuracy is an indicator for the comprehension level, i.e. the higher the accuracy of the transcription the higher the comprehension.

We used Amazon Mechanical Turk to run the user study. We took several precautions to guarantee the quality of the data (burch, 2009). We restricted the workers to those who have more than 95% approval rate for all their previous work and who live in the United States (since we are targeting English speakers). We also assigned the same audio clip to 10 different workers and took the average edit distance of the 10 transcripts for each audio clip.

The differences in the transcription accuracy for the original and the transformed paragraphs were statically significant at the 0.05 level according to a 2-tailed paired t-test. The overall average edit distance was 0.69 for the 50 transformed paragraphs and 0.56 for the original article paragraphs. This result indicates that the change in style has an impact on the comprehension of the delivered information as measured by the accuracy of the transcriptions.

7 Conclusions and Future Work

In this paper, we presented the progress on an ongoing research on writing style transformation from text style to audio style. We motivated the topic and emphasized its importance. We surveyed the linguistics and journalism literatures for the differences in writing style for different media. We formalized the problem by suggesting a number of linguistic features and showing their validity in distinguishing between the two styles of interest, text vs audio. We also conducted a user study to show the impact of style transformation on comprehension and the overall user experience.

The next step in this work would be to build a style transformation system that uses the features discussed in this paper as the bases for determining when, where, and how to do the style transformation.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346.

Shlomo Argamon, Paul Chase, Sushant Dhawle, Sobhan Raj, Hota Navendu, and Garg Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science. ((In press)) Baayen*, 7:91–109.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Chris Callison burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazons mechanical turk.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Jrgen Esser. 1993. *English linguistic stylistics*. Niemeyer.

Jrgen Esser. 2000. Medium-transferability and presentation structure in speech and writing. *Journal of Pragmatics*, 32.

Irving E. Fang. 1991. *Writing Style Differences in Newspaper, Radio, and Television News*. Monograph Ser Vol, 1. University of Minnesota. Center for Interdisciplinary Studies of Writing.

Robert Gunning. 1952. *The technique of clear writing*.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422. 10.1023/A:1012487302797.

Michael Alexander Kirkwood Halliday. 1985. *Spoken and Written Language*. Deakin University Press.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*, 36:341–387.

Frederick Mosteller and David L. Wallace. 1964. *Inference and disputed authorship : the Federalist / [by] Frederick Mosteller [and] David L. Wallace*. Addison-Wesley, Reading, Mass. :.

Richard Power and Donia Scot. 2005. Automatic generation of large-scale paraphrase.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

Jonathan Schler and Shlomo Argamon. 2009. Computational methods in authorship attribution.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 313–318, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis T. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26:471–495.

Steve Whittaker, Julia Hirschberg, and Christine H. Nakatani. 1998. Play it again: a study of the factors underlying speech browsing behavior. In *CHI '98: CHI 98 conference summary on Human factors in computing systems*, pages 247–248, New York, NY, USA. ACM.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 834–842, Stroudsburg, PA, USA. Association for Computational Linguistics.