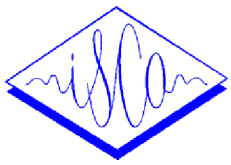
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Towards Interpretation of Creakiness in Switchboard

Xiaodan Zhuang and Mark Hasegawa-Johnson

Department of Electrical & Computer Engineering
Beckman Institute of Science & Technology
University of Illinois Urbana-Champaign
{xzhuang2,jhasegaw}@uiuc.edu

Abstract

This paper adopts Latent Semantic Analysis (LSA) for long-term analysis of voice quality, in particular creakiness. Each automatically labeled creaky instance (word) is modeled as a $document$ and different prosodic and syntactic cues as $terms$. This framework attempts to automatically identify the most salient correlates, or $latent\ factors$, of creakiness, and further assign each creaky instance (word) to one of the latent factors. The algorithm implemented in this study identifies at least two correlates of creakiness in Switchboard: (1) particles, coordinating conjunctions in repair/repeat locations, and filled pauses; (2) starts of various sentence/clause structures, such as Wh-adverb phrases, sentences and asides with sentence restarts at repair/repeat locations. Such automatic long-term voice quality analysis could pave the way for better incorporating voice quality in speech recognition, among other speech applications.

10.21437/SpeechProsody.2008-8

1. Introduction

The acoustic source of speech sounds, especially voiced speech sounds, is defined as the airflow through the glottis. Quasi-periodic vibration of the vocal folds results in a volume velocity waveform. Voice quality refers to the quality of sound produced with a particular setting of the vocal folds, and includes breathy, creaky and modal voices. In particular, creaky voice, associated with slow and irregular vocal pulses, has been identified to bear some linguistic functions in American English, including encoding allophonic variation [1], encoding junctures [2, 3, 4, 1] and signaling tiredness or boredom.

Given the functions of creakiness in American English, it is possible to improve automatic speech recognition by modeling such voice quality variation. In [5], creakiness is explicitly modeled in triphone acoustic models of a speech recognizer, achieving word accuracy improvement. Information about voice quality could further be used to favor recognition hypotheses in which higher-level, i.e. syntactic and prosodic, structures are consistent with the observed creakiness. In this way, voice quality constitutes a new channel of information to guide recognition of American English utterances. This calls for identifying correlates of the creaky instances so that an appropriate constraint could be applied.

On the other hand, while progress has been made using expert knowledge to identify some functions of creakiness [2, 3, 4, 1], there does not yet exist a comprehensive list of the syntactic, prosodic and pragmetic correlates of creaky voice. Indeed, it is not clear that an inviolably complete list could be constructed: creakiness can be generated at will, therefore novel meanings of creakiness might be generated, at any time, by any community of speakers. What we can do, however, is to semi-automatically catolog the uses of creakiness in a fixed database, e.g. the Switchboard corpus of conversational American English. An automatic data-driven approach to identify the correlates, or $latent\ factors$, of creaky instances is of particular interest for various reasons. First, human analysis is very expensive, particularly when various correlates come into interplay. Second, a data-driven approach might help discover new correlates in large corpora. Third, automatic analysis makes it possible to apply the analysis results in engineering applications such as speech recognition and understanding.

We propose using Latent Semantic Analysis (LSA), motivated by its success in modeling latent factors linking different documents and individual words in natural language processing applications. In this study, LSA is applied on an $instance-term\ matrix$, where each column corresponds to one creaky $instance$ and each row corresponds to one particular $term$ or feature of interest. The creaky instances in the original high-dimensional $term\ space$ are mapped to the low-dimensional $latent\ space$, which hopefully captures the major correlates of creakiness. Thus the approach not only identifies the major correlates, but labels each creakiness instance in the corpus with one of these correlates.

2. Creakiness & Objective Labeling

2.1. Voice Quality Categories

Ladefoged [6] suggests that types of voice quality, or phonation types, be defined in terms of the aperture between the arytenoid cartilages in the larynx. The degree of aperture between the vocal folds plays a role in producing voice qualities such as modal, breathy, and creaky voices. Modal voice refers to the phonation of speech sounds produced with regular vibrations of the vocal folds, thus with relatively well-defined pitch pulses. Breathy voice is characterized by vocal cords that are fairly abducted and have little longitudinal tension. The abduction and lesser tension allow some turbulence of airflow to flow through the glottis. Creaky voice is typically associated with vocal folds that are tightly adducted but open enough along a portion of their length to allow for voicing. Due to the tight adduction, the creaky voice typically reveals slow and irregular vocal pulses in the spectrogram, where the vocal pulses are farther apart from each other compared to those of modal and breathy voices.

Functions of voice quality include encoding lexical contrast [7, 8, 9], encoding allophonic variation [1], signaling speaker's emotional or attitudinal status [10], signaling sociolinguistic or extra-linguistic indices [11], and marking junctures

This research was funded by NSF grant IIS 04-14117. All results and conclusions are those of the authors, and do not necessarily represent the opinion of the National Science Foundation.

[2, 12, 4]. The utilization of the voice quality functions could be language-dependent.

Among the categories of voice quality, creaky voice has been found to signal linguistic information in American English. Creakiness in American English encodes allophonic variation [1] and is further related to prosodic structure as a frequent correlate of word, syntactic, or prosodic boundaries [2, 3, 4, 1].

2.2. Objective Labeling

Acoustic cues obtained from voice source analysis, in particular spectrum analysis, have been found more reliable for voice quality identification than fundamental frequency (F0) or intensity alone. Nł Chasaide and Gobl [13] characterized creaky phonation as having slow and irregular glottal pulses, in addition to low F0. Specifically, they state that significant spectral cues to creaky phonation are (1) $A1$ (amplitude of the strongest harmonic of the first formant) much higher than $H1$ (amplitude of the first harmonic) , and (2) $H2$ (amplitude of the second harmonic) higher than $H1$. Yoon et al [14] also used spectral features including $H1$-$H2$ to classify subjective voice quality with 75% accuracy.

This work adopts the voice quality decision approach in [5], which detects creaky voice quality based on acoustic cues, independent of higher-level linguistic context. Interactively-determined thresholds are used to divide the two-dimensional feature space of temporal mean autocorrelation (Rx) and amplitude difference between the second and first harmonics ($H1$-$H2$) into a set of voice-quality-related objective categories. For each 10ms frame, the "voiceless" category includes all frames for which no pitch can be detected. The "creaky" category includes all frames for which $H1$-$H2 <$ -15 dB, or for which $H1$-$H2 < 0$ and $Rx < 0.7$. All other frames are assigned to an objective category called "modal." Using a word transcription without boundary information and a dictionary spelling words into phone sequences, we use a speech recognizer to obtain the time-aligned phone transcription of Switchboard. Within the boundaries of each sonorant phone, if more frames indicate creaky category than any other category, the phone is labeled as creaky. Only sonorants, not obstruents such as stops and fricatives, are eligible to be assigned the creaky label. Any word having at least one creaky phone is labeled as a creaky word.

3. Syntactic & Prosodic Tags

Much work has shown supersegmental correlates of creakiness in English. Kushan and Slifka [2] report that 5% of their 1331 hand-labeled irregular tokens in a subset of TIMIT database occur at syllable boundaries, and 78% of the tokens at word boundaries. Laver [12] states that creaky voice with a concomitant low falling intonation may be used by speakers of English as a marker for turn taking. Dilley et al. [3] show, through the analysis of prosodically labeled American English, that phrasal boundaries of intermediate and intonational phrases influence glottalization of word-initial vowels. Redi and Shattuck-Hufnagel [4] further demonstrate that glottalization is more likely to be observed on words at the ends of utterances than on words at the ends of utterance-medial intonational phrases, and that the glottalization is more likely to be observed on boundaries of full intonational phrases than on boundaries of intermediate phrases.

To make long-term analysis of voice quality feasible, it's preferable to have access to both prosodic and syntactic tags on the corpus. Many of the prosodic tags, however, are not available on spontaneous speech corpora such as Switchboard, and are extremely expensive to manually generate. On the other hand, many readily available tags, such as syntactic tags and speech disfluency tags are tightly related to prosody. It is possible to capture the prosodic variation using syntactic and disfluency labeling [15].

Therefore, the $terms$, in the Latent Semantic Analysis terminology, could include the disfluency tags, the syntactic tags, silence, word fragment, speaker, utterance, non-speech sounds, etc. In this study, we only use the disfluency tags and the syntactic tags.

3.1. Disfluency Tags

Disfluencies in human speech are among the characteristics that differentiate spontaneous speech from read speech. Repetitions, filled pauses and deletions are the most frequent across various spontaneous speech corpora, defined by how human subjects would process them. It has been shown that prosodic cues could be used to predict disfluencies [16], indicating the strong correlation between prosody and speech disfluency.

Disfluency tags used in this study include Penn Treebank disfluency tags. We adopt the notation of the disfluency interval tags and the non-sentence element tags in a disfluency transcription aligned to Switchboard ms98 word transcription by [17].

3.2. Syntactic Tags

The syntactic Tags could include the part of speech (POS) tags and bracketing syntactic tags in Penn Treebank [18]. Some examples of syntactic tags are as follows:

- ADJP - Adjective phrase
- DT - Determiner
- JJS - Adjective superlative
- UH - Interjection
- WDT - Wh-determiner
- WHADVP - Wh-adverb phrase
- WHNP - Wh-noun phrase
- WHPP - Wh-prepositional phrase

4. Voice Quality Analysis Using Latent Semantic Analysis

4.1. Latent Semantic Analysis

Latent semantic analysis (LSA) was introduced into information retrieval [19] to tackle the problem that lexical matching at term (word) level is inaccurate owing to polysemy and synonomy. In latent semantic analysis, a large term by document matrix is constructed from raw text and then decomposed using singular value decomposition into a set of (much fewer than the number of word types) orthogonal factors from which the original matrix can be approximated by linear combination.

A $term\ by\ document\ matrix\ W$ is a compact summary of a set of $documents\ \{d_1, d_2, ..., d_d\}$, corresponding to the columns, with the vocabulary, i.e. a set of $terms\{t_1, t_2, ..., t_t\}$, corresponding to the rows. Each element w_{ij} of the matrix is a normalized count of term t_i appearing in document d_j. An arbitrary rectangle matrix with different entities on the rows and columns, such as the term by document matrix W, can be decomposed and approximated by three special matrices as

$W = TSD^T$. Matrix T and Matrix D are composed of row singular vectors, each corresponding to one term or one document respectively. Matrix S have non-zero elements only on its diagonal, encoding the strength of the latent factors.

To map any new document, called $query$ in information retrieval terminology, into the latent space, simply construct a column vector X_q in the same way as any column vector in matrix W: the column vector X_q specifies the counts of each term in the query. Then $X_q^T T$ maps X_q into the latent space.

4.2. Modeling Creaky Instances

To apply LSA to long-term voice quality analysis, in particular creakiness analysis, we formulate the problem as follows, around the notion of $instance - term\ matrix$, as an analogy to the $term\ by\ document\ matrix$ in information retrieval.

Each creaky instance in the corpus is modeled as a document. Such instance could be either a phone labeled as creaky, a syllable including a creaky phone or a word including a creaky phone. If the previous two approaches are adopted, we might not only capture prosodic and syntactic correlates of creakiness, but also its phonological correlates. However, that would also lead to more severe data sparseness problem. Therefore, this study adopts the third approach, i.e. each creaky word modeled as a document.

All tags assigned to any creaky word are modeled as the set of terms. Given that some tags might span more than one word, the actual tags used are in the form of "ST_TAG" or "EN_TAG", where "ST_" denotes that this word is the starting point of the region tagged with "TAG", and "EN_" denotes the ending point. This increases the number of tags to twice of that seen in Section 3 and enables a finer representation of the syntactic and disfluency features.

We expect that the latent factors derived from the term by document matrix defined above would explain in some sense the correlates of creakiness in the corpus. It could be hard to directly interpret the meanings of these latent factors for at least the following reasons. First, though latent factors were demonstrated to capture the underlying topics in multi-topic corpora, they could be hard to interpret in other applications, such as information retrieval [19]. Second, there isn't yet a comprehensive list of the correlates of creakiness in spontaneous American English. While the latent factors might correspond to some known reasons for a word being creaky, they might also reveal something beyond the known linguistic theories, in the data-driven perspective, thanks to the applicability of approaches proposed in this study on a huge speech corpus with syntactic and disfluency tags.

By constructing a query vector for a term, or a tag, by setting count one only for the element corresponding to that particular term, the coordinates of that term in the latent semantic space could be calculated according to Section 4.1. These coordinates reveal the "strength" of that term on each of the latent factors, which can be used to assign the term to the latent factor with the most strength.

By modeling each creaky word as a query with count one only for the terms assigned to that word, it gets mapped to the latent space in a similar way as above. More specifically, we could label each creaky word with the latent factor with the most strength, leading to a hopefully more compact, in the prosodic or syntactic sense, label set than the single noisy "creaky" label.

5. Experiments

5.1. Experiment Setup

This experiment is carried out on a subset of about 3,300 utterances in Switchboard corpus, including about 29,000 words, among which about 7,000 are labeled as "creaky" by the objective labeling presented in Section 2.2. The $terms$ used in this experiment are the syntactic terms and the disfluency terms described in Section 3, combined with "ST_" and "EN_" in Section 4.2. Figure 1 illustrates that a creaky word is labeled via the objective labeling scheme, and further labeled with an appropriate latent factor. (The cirles indicate creaky phone or word.)

Figure 1: Latent factors assigned to creaky words

5.2. Interpreting Latent Structure

Table 1 shows the terms assigned to each latent factor in singular value decomposition with five latent factor approximation. The terms are assigned to the latent factor on which its projection is highest. Only the most salient tags for each latent factor are listed in this table, with saliency defined as the difference between the projection on the latent factor with highest projection and that with the second.

Table 1: Most salient tags assigned to five latent factors

Latent 1	
Latent 2	ST_ADJP, EN_JJS
Latent 3	ST_WDT, EN_RP, EDT/C, ST_PDT, ST_UH, EN_RBS, ST_PRN
Latent 4	EN_S1, EDT/A, ST_WHADJP, ST_VP, ST_S1, EN_POS, EN_DT
Latent 5	ST_WHNP, ST_WHPP

Expectedly noisy as it is, we may attempt to interpret the latent factors. The latent factor 1 has no tags assigned to it because this first latent factor is usually associated with the general mean of all tags as is also shown in some other applications. Latent factor 2 has terms or tags such as the starting word of an adjective phrase and the ending word of a superlative adjective. Latent factor 3 has, among others, Wh-determiner, particle, coordinating conjunction in a repair/repeat location, predeterminer and filled pause. Latent factor 4 has starts of various sentence/clause structures, such as Wh-adverb phrase, aside with sentence restart at repair/repeat location, verb phrase and sentence. Latent factor 5 has the starts of Wh-noun phrase and Wh-prepositional phrase.

5.3. Labeling Words with Latent Factor

By mapping all creaky words in the dataset into the latent space, we can assign words to the latent factor on which it has the highest projection. Only words having much higher energy on one

latent factor than the others are counted. This result is presented in Table 2.

Table 2: Creaky words assigned to latent factors

Latent 1	Latent 2	Latent 3	Latent 4	Latent 5
1734	457	1212	2922	263

The large number of creaky words assigned to latent factor 1 presents a significant difference with the tag assignment results. This might be explained by the fact that these creaky word instances approximate themselves closer to the general mean of the term by document matrix, represented by latent factor 1, than to any of the projections on the other latent factors.

5.4. Most Frequent Tags in Each Latent Factor

Among creaky words labeled by a particular latent factor other than latent factor 1, the most frequent tags are presented in Table 3. This result is very similar to Table 1.

Table 3: Most frequent tags assigned to four latent factors

Latent 2	ST_ADJP, EN_JJS
Latent 3	ST_PRN, ST_WDT, ST_UH, EDT/C, ST_PDT, EN_RP, EN_RBS
Latent 4	ST_WHADJP, ST_VP, EDT/A, EN_DT, ST_S1, EN_S1, EN_POS
Latent 5	ST_WHNP, ST_WHPP

6. Conclusion & Discussion

This work adopts a widely-used technique in information retrieval, Latent Semantic Analysis (LSA), for long-term analysis of voice quality, in particular creakiness. Each creaky word instance is modeled as a document and different prosodic and syntactic tags as terms. This framework attempts to automatically identify the most salient correlates, i.e. latent factors, of creaky voice, and further assign each creaky instance to one of the latent factors.

The results of long-term voice quality analysis using LSA is noisy as expected. Looking at the tags assigned to the five latent factors, they do somehow correspond to the long-term functions creakiness is believed to have. The LSA framework implemented in this work successfully identifies at least two commonly accepted correlates of creakiness: 1) particle, coordinating conjunction in a repair/repeat location, and filled pause; 2) starts of various sentence/clause structures, such as Wh-adverb phrase, aside with sentence restart at repair/repeat location, and sentence. Such automatic long-term voice quality analysis could pave the way for incorporating voice quality in speech recognition beyond local acoustic modeling.

The goal of this paper is not to propose the latent factors found in this particular experiment as a universalizable listing of the correlates of creaky voicing. The details of this listing vary depending on the corpus studied, its annotation and details of analysis including data normalization. Instead, the goal of this paper is to propose the LSA framework as a possible approach in acoustic phonetic corpus study. We suspect that by more carefully constructing the tag set, for example, by including other tags such as speaker information and dialogue topic, by grouping part-of-speech sets into content word and function word, by trying different numbers of latent factors, or by looking at the counts of some particular tags over a window of several words rather than one single word, LSA has the potential to capture more interesting phenomena. It would also be interesting to see whether these latent factors can serve as helpful labels for applications such as speech recognition and understanding.

7. References

[1] M. A. Epstein, *Voice Quality and Prosody in English*, Ph.D. dissertation, University of California Los-Angles, California, 2002.

[2] S. Kushan and J. Slifka, "Is irregular phonation a reliable cue towards the segmentation of continuous speech in american english," in *ICSA International Conference on Speech Prosody, Dresden, Germany*, 2006.

[3] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalizationof word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, pp. 423–444, 1996.

[4] L. Redi and S. Shattuck-Hufnagel, "Variation in the rate of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407–427, 2001.

[5] Tae-Jin Yoon, Xiaodan Zhuang, Jennifer Cole, and Mark Hasegawa-Johnson, "Voice quality dependent speech recognition," *Language and Linguistics*, 2007, In preparation.

[6] Peter Ladefoged, *Preliminaries to Linguistic Phonetics*, University of Chicago Press, Chicago, 1971.

[7] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*, Blackwell, Oxford, UK, 1997.

[8] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, pp. 383–406, 2001.

[9] E. Fischer-Jorgensen, "Phonetic analysis of breathy (murmured) vowels," *Indian Linguistics*, vol. 28, pp. 71–139, 1967.

[10] C. Gobl, *The Voice Source in Speech Communication: Production and Perception Experiments Involving Inverse Filtering and Synthesis*, Ph.D. dissertation, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 2003.

[11] J.H. Esling, *Voice quality in Edinburgh: a sociolinguistic and phonetic study*, Ph.D. dissertation, University of Edinburgh, U.K., 1978.

[12] J. Laver, *The Phonetic Description of Voice Quality*, Cambridge University Press, 1980.

[13] A. Ni Chasaide and C. Gobl, "Voice source variation," in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds., pp. 1–11. Blackwell, Oxford, UK, 1997.

[14] Tae-Jin Yoon, Jennifer Cole, Mark Hasegawa-Johnson, and Chilin Shih, "Acoustic correlates of nonmodal phonation in telephone speech," *Journal of the Acoustical Society of America*, 2007.

[15] R. Kompe, *Prosody in speech understanding systems*, Springer, Berlin, 1997.

[16] E.E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proceedings of EUROSPEECH, Rhodes, Greece*, 1997, pp. 2383–2386.

[17] K. Gorman, "Time-aligned switchboard disfluency corpus," Tech. Rep., Department of Linguistics, University of Illinois Urbana-Champaign, Urbana, IL, 2005.

[18] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[19] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.