Voice Quality Dependent Speech Recognition

Tae-Jin Yoon, Xiaodan Zhuang, Jennifer Cole, & Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign, USA

Abstract

Voice quality conveys both linguistic and paralinguistic information, and can be distinguished by acoustic source characteristics. We label objective voice quality categories based on the spectral and temporal structure of speech sounds, specifically the harmonic structure (H1-H2) and the mean autocorrelation ratio of each phone. Results from a classification experiment using a Support Vector Machine (SVM) classifier show that allophones that differ from each other regarding voice quality can be classified as distinct using input features in speech recognition. Among different possible ways to incorporate voice quality information in speech recognition, we demonstrate that by explicitly modeling voice quality variance in the acoustic phone models using hidden Markov modeling, we can improve word recognition accuracy.

Keywords: ASR, Voice quality, H1-H2, Autocorrelation ratio, SVM, HMM.

1. Introduction

The acoustic source of speech sounds, especially the source of voiced speech sounds, is defined as the airflow through the glottis. Quasi-periodic vibration of the vocal folds results in a volume velocity waveform. The source signal is modulated in the vocal tract, which functions as a resonator or a filter (Fant 1960). The term "voice quality" refers to the quality of sound produced with a particular setting of the vocal folds, and includes breathy, creaky and modal voices. Voice quality provides information at multiple levels of linguistic organization, and manifests itself through acoustic cues including F0, and information in spectral and temporal structures. If we can reliably extract acoustic features that differentiate phones on the basis of voice quality, then voice quality differences can be modeled in an Automatic Speech Recognition system (ASR), improving recognition performance.

Fundamental frequency (F0) and harmonic structure are acoustic parameters that signal voice quality. Particularly, they are shown to be important factors in encoding lexical contrast and allophonic variation related to laryngeal features (Maddieson and Hess 1987; Gordon and Ladefoged 2001). For example, Maddieson and Hess (1987) observe significantly higher F0 for tense vowels in languages that distinguish three phonation types (tense, lax, and modal) with varying voice quality (Jingpho, Lahu and Yi). However, F0 is not always a reliable indicator of voice quality. Studies of English have failed to show a strong correlation between any glottal parameters and F0 (Epstein 2002). On the other hand, information obtained from harmonic structure has been shown to be more reliable for the discrimination of non-modal from modal phonation. For example, Gordon and Ladefoged (2001) describe the characteristics of creaky phonation as producing non-periodic glottal pulses, lower power, lower spectral slope, and low F0. Among these acoustic features, they report that spectral slope is the most important feature for discrimination among different phonation types.

The observation of voice quality differences, even non-phonemic differences as in English, raises a research question: Will the incorporation of voice quality into a speech

recognition system result in improved performance? We hypothesize that the spectral characteristics of phones produced with creaky voice are so different from those produced with modal voice that direct modeling of voice quality will result in improved word recognition accuracy. We test that hypothesis in the present study by labeling the voice quality of spontaneous connected speech using both harmonic structure (a spectral measure) and mean autocorrelation ratio (a temporal measure), which have been identified to be reliable indicators of voice quality.

Speech is usually parameterized as perceptual linear prediction (PLP) coefficients in speech recognition systems, to reflect human auditory characteristics. An important question is whether these parameters used in ASR also reflect voice quality variation. We answer this question by showing that the phone-level voice quality labels automatically generated according to a spectral measure taken from harmonic structure and a temporal measure of mean autocorrelation ratio are predictive of their PLP coefficients. We further show that a PLP-coefficients-based automatic speech recognizer that incorporates voice quality information in the acoustic models performs better than a complexity-matched baseline system that does not consider the voice quality distinction.

The paper is organized as follows. Section 2 illustrates linguistic and paralinguistic functions of voice quality (subsection 2.1) and presents acoustic cues for the voice quality identification (subsection 2.2). Section 3 introduces our method of voice quality decision on the corpus of telephone conversation speech. Section 4 reports a classification result that shows the voice quality distinctions are reflected in PLP coefficients. Section 5 presents an HMM-based speech recognition system that incorporates voice quality knowledge. Section 6 compares the performance of the voice quality dependent recognizer against a baseline system that doesn't distinguish different voice qualities. Section 7 concludes the paper with discussion of the source of the ASR improvement in the increased precision of the phone models that are specified for different voice qualities.

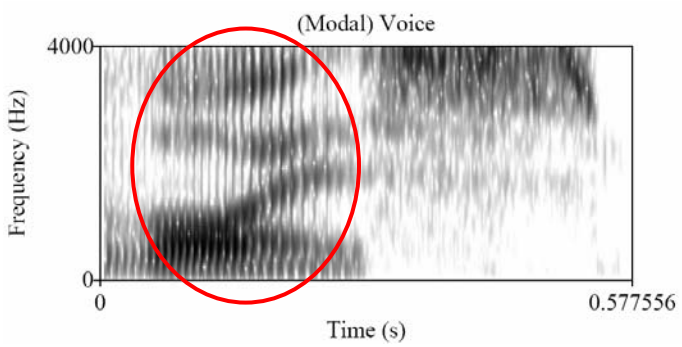
2. Voice Quality

Among numerous types of voice quality (e.g., see Gerratt and Kreiman 2001), the most frequently utilized cross-linguistically are modal, creaky, and breathy voices. In this section, we briefly illustrate the characteristic of voice qualities, and present uses and functions of voice quality (subsection 2.1) and acoustic correlates of the types of voice quality (subsection 2.2).

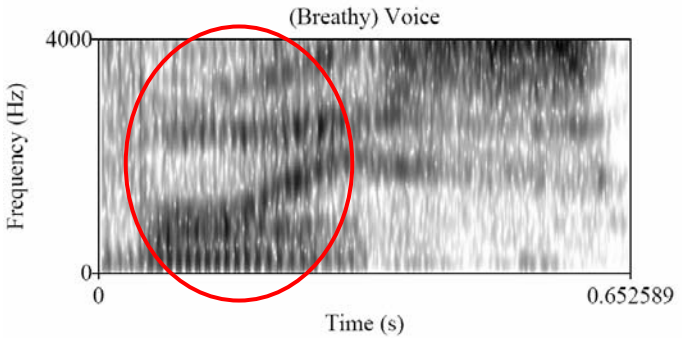
Ladefoged (1971) suggests that types of voice quality, or phonation types, be defined in terms of the aperture between the arytenoid cartilages in the larynx. The arytenoid cartilages are a pair of small three-sided components in the larynx. The vocal folds are attached to these cartilages. The degree of aperture between the arytenoid cartilages, hence between the vocal folds, plays a role in producing voice qualities such as modal, breathy, and creaky voices. Modal voice, as is illustrated in Figure (1a), refers to the phonation of speech sounds produced with regular vibrations of the vocal folds. The modal voice has relatively well-defined pitch pulses. In Figure (1a), relatively well defined striations in the formants are visible in the region where the vowel [oi] in the word 'voice' is uttered. Breathy phonation, as is shown in Figure (1b), is characterized by vocal cords that are fairly abducted (relative to modal and creaky voice) and have little longitudinal tension. The abduction and lesser tension allow some turbulence of airflow to flow through the glottis. In Figure (1b), turbulent noise is present across the frequency range. Creaky phonation, as in Figure (1c), 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, as in Figure (1c), where the vocal pulses are farther apart from each other compared to those of modal and breathy voices in Figures (1a-b)[1].

(a)



(b)



(c)

Figure 1: Spectrograms of the same word "voice" that are produced with different phonation qualities. From top to bottom, the word "voice" is produced with (a) modal voice, (b) breathy voice, and (c) creaky voice, respectively. The circles in the above figures indicate regions where different types of voice quality are observed.

2.1 Functions of voice quality

Functions of voice quality include the encoding of lexical contrasts, encoding of allophonic

[1] The sound files are taken from http://www.ims.uni-tuttgart.de/phonetik/EGG/

variation, signaling of speaker's emotional or attitudinal status, and socio-linguistic or extra-linguistic indices. The utilization of the voice quality function is language-dependent.

The use of voice quality to encode lexical contrasts is fairly common in Southeast Asian, South African and Native American Languages. For example, the presence or absence of creakiness on the vowel *a* in "já" signals difference in meaning in Jalapa Mazatec such that "já̰" produced with creakiness means "he carries" whereas "já" produced without creakiness means "tree" (Ladefoged and Maddieson 1997; Gordon and Ladefoged 2001). Gujarati speakers need breathy voice or murmured voice to distinguish the word /bə̤r/ produced with murmured voice "outside" from the word /bar/ "twelve" (Fischer-Jørgensen 1967; Bickley 1982; Gordon and Ladefoged 2001)[2].

Voice quality is also commonly used to encode allophonic variation in certain contexts. That is, many languages use non-modal phonation in creaky or breathy voice as variants of modal voice in certain contexts. For example, voiceless stop /t/ in American English is often realized as glottal stop [ʔ]. The spectrogram in Figure 2 illustrates that the final /t/ in the word "cat" is produced with glottal stop [ʔ], with anticipatory non-modal phonation on the preceding vowel.

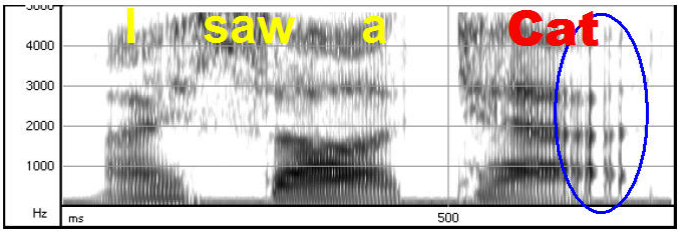


Figure 2: An allophonic realization of the voiceless stop /t/ as a glottal stop [ʔ] (Figure taken from Epstein 2002)

A particular voice quality is more likely to be associated with specific tones in tonal languages. Huffman (1987) observes that one of the seven tones in Hmong (a Sino-Tibetan language) is more likely to occur with a breathy voice quality. Jianfen and Maddieson (1989) describe that the yang tone in the Wu dialect of Chinese differ from the yin tone in that the yang tone is associated with the breathy voice.

Voice quality can function as a marker for juncture. For example, creaky voice can be used to mark syllable, word, phrase, and utterance boundaries. Kushan and Slifka (2006) 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. For example, creakiness is observed at the end of a word boundary in Figure 3.

[2] For names of languages with different types of phonation contrasts, see Gordon and Ladefoged (2001).

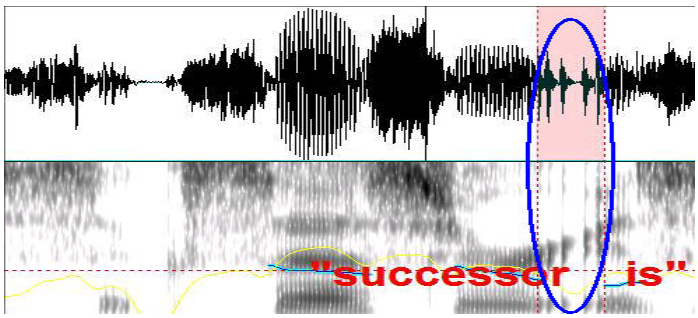


Figure 3: An example of the occurrence of creakiness at a word boundary. Creakiness is used in the realization of the rhotic "r" at the end of the word "successor."

Fant and Kruckenberg (1989) demonstrate that creaky voice is used as a phrase boundary marker for speakers of Swedish. Laver (1980) states that creaky voice with a concomitant low falling intonation may used by speakers of English as a marker for turn taking. Dilley et al. (1996) show, through the analysis of a prosodically labeled speech corpus of American English, that phrasal boundaries of intermediate and intonational phrases influence glottalization of word-initial vowels. Redi and Shattuck-Hufnagel (2001) 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.

In addition to the linguistically determined variation discussed above, there are paralinguistic functions in the use of voice qualities. Modulation of voice quality can be used to convey the speaker's emotion and attitude to the listener. For example, creaky voice signals tiredness or boredom, at least in American English. It should be noted that the use of voice quality and its relation to emotional or attitudinal aspects do not seem to be universal. For many speakers of Swedish, creaky voice is an affectively unmarked quality, whereas the same voice quality is used in Tzeltal (a Mayan language) to express commiseration or complaint (Gobl 2003) and it is use in Slovene to express indecisiveness or uncertainty[3]. In addition, breathy voice is associated with intimacy in many languages. The affect of intimacy is typically regarded to be a marker for female speakers rather than a marker for male speakers. For example, Gobl (2003) states that "gender-dependent differences, particularly increased breathiness for female speakers, have been observed in languages," including English.

Finally, it has been observed that voice quality may also have a sociolinguistic dimension serving to differentiate among social groups. Within a particular dialect, voice quality features may signal social subgroups. Esling (1978, quoted in Gobl 2003) states that "in Edinburgh English, a greater incidence of creaky voice is associated with a higher social status, whereas whispery and harsh qualities are linked to a lower social status."

Among the categories of voice quality, creaky voice has been recurrently reported to play a role in American English in signaling linguistic information, even though the function of creakiness in American English is not phonemic. Creakiness in American English is related to prosodic structure as a frequent correlate of word, syntactic, or prosodic boundaries (Kushan and Slifka 2006; Dilley et al. 1996; Redi and Shattuck-Hufnagel 2001; Epstein 2002). Given the linguistic function of creakiness in American English, it is possible to use voice quality to facilitate automatic speech recognition. Information about voice quality can be used to decide

[3] http://www2.ku.edu/~slavic/sj-sls/jurgec_eng.pdf

between candidate analyses of an utterance by favoring analyses in which the syntactic and higher-level structures are consistent with the observed voice quality of a target word. In this way, voice quality constitutes a new channel of information to guide phrase-level analysis. An even more basic benefit of voice quality information is also possible: Voice quality effects condition substantial variation in the acoustic realization of a word or phone. Modeling that variation offers the possibility of improved accuracy in word or phone recognition. The next section details a method for reliably detecting creaky voice quality based on acoustic cues, independent of higher-level linguistic context, for the purpose of modeling creaky voice for speech recognition.

2.2 Acoustic correlates of voice quality

Acoustic cues obtained from voice source analysis have been identified to be more reliable for voice quality identification than F0 or intensity alone. But analytic studies have largely focused on the more measurable parameters of F0 and intensity (cf. Gordon and Ladefoged 2001; Gobl 2003). This can be attributed to the methodological difficulties in voice source analysis with features other than F0 or intensity. For example, segments with both breathy and creaky voices have been shown to have reduced intensity characteristics. In certain languages such as Chong (Thongkum 1987) and Hupa (Gordon 1996), it has been observed that phones produced with creakiness trigger a reduction in intensity relative to the intensity observed in phones produced with modal phonation. However, the intensity measurement is subject to many external factors such as location of the microphone and background noise, and internal factors such as the speaker's loudness level. Slow and irregular vibration of the vocal folds characterizes creaky voice, resulting in low F0. However, F0 is not always a reliable indicator of voice quality. Studies of English have failed to show a strong correlation between any glottal parameters and F0 (Epstein 2002).

Information obtained from spectral structure is more reliable for the voice quality identification. Ní Chasaide and Gobl (1997) characterize 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 i) A1 (i.e., amplitude of the strongest harmonic of the first formant) much higher than H1 (i.e., amplitude of the first harmonic)[4], and ii) H2 (i.e., amplitude of the second harmonic) higher than H1.[5] (See Figure 4 for an illustration.) Fisher-Jørgensen (1967) conducted a discrimination experiment between modal vowels and breathy vowels with Gujarati listeners using naturally produced Gujarati stimuli. The listeners were able to distinguish breathy vowels from modal ones in cases where the amplitude of the first harmonic dominates the spectral envelope. She observed that other cues such as F0 and duration had little importance in the task. Pierrehumbert (1989) investigated the interaction of prosodically prominent events such as pitch accents and voice source variables. In general, the glottal pulse for high toned pitch

[4] Relative contribution of the A1 to the creaky voice is related to the increased bandwidth of the first formant. Hanson et al. (2001) states that "if the first-formant bandwidth (B1) increases, the amplitude A1 of the first-formant peak in the spectrum is expected to decrease. Therefore, the relative amplitude of the first harmonic and the first-formant peak (H1-A1, in dB) is selected as an indicator of B1" Thus, the relative difference between H1-A1 is relevant to the discrimination between creaky voice and modal voice.

[5] Some researchers use H1 and H2 to refer to individual harmonics, not to the amplitudes of thereof. In this paper, H1 and H2 refer to the amplitude of each harmonic, i.e., first and second harmonics, respectively.

accents has a greater open quotient than for low toned pitch accents. The open quotient (OQ) is defined as the ratio of the time in which the vocal folds are open to the total length of the glottal cycle. But it is also occasionally observed that while higher voice level as measured by intensity results in a higher F0, the higher voice level corresponds to a reduced OQ. This implies again that the F0 and cues from voice source are largely independent of each other, and the open quotient, which is related to the harmonic structures of H1 and H2, provides a more reliable cue for the identification of non-modal phonation.

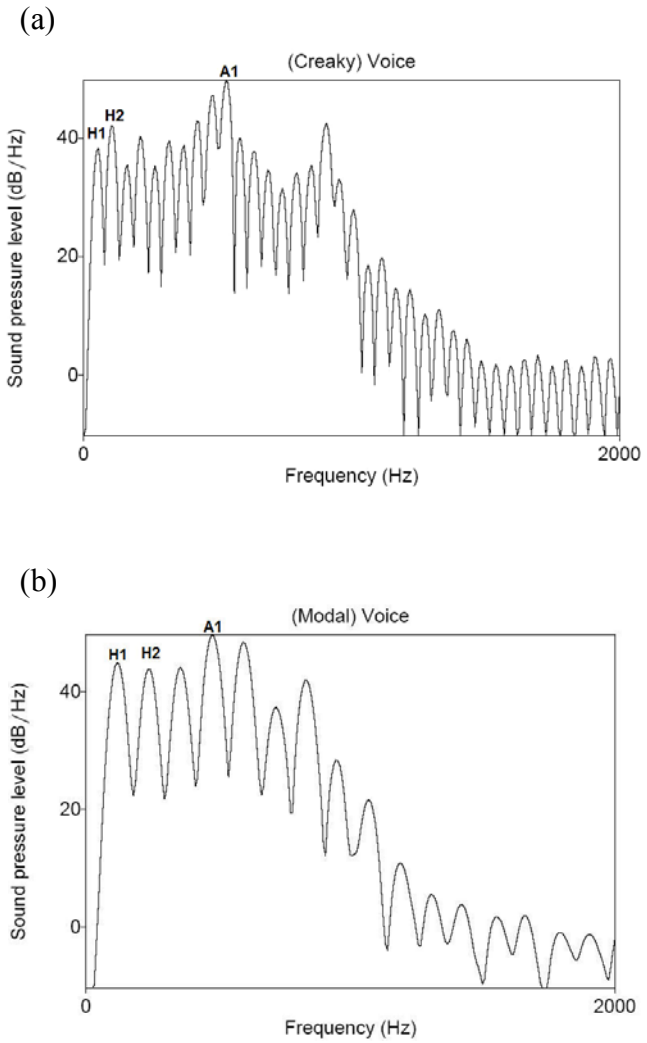


Figure 4: Spectral slices taken from the vowel 'oi' in the word "voice" (a) when the vowel is produced with creaky voice, and (b) when the vowel is produced with modal voice. In (a), both H2 and A1 are relatively higher than H1. In (b), H2 is approximately the same as H1, and A1 is relatively higher than H1.

H1 and H2 are related to the open quotient (OQ) (Fant 1997; Hanson and Chuang 1999; Hanson et al. 2001). The numerical relationship between H1-H2 and OQ is reported in Fant (1997) as in (1):[6]

[6] In the literature, H1*-H2* is sometimes used instead of H1-H2. H1*-H2* is a modification of H1-H2

$$H1 - H2 = -6 + 0.27\exp(5.5 \times OQ) \quad (1)$$

In creaky voicing, the vocal folds are held tightly together (though often with low internal tension), resulting in a low OQ. That is, the more the amplitude of the second harmonic relative to that of the first harmonic, the lesser is OQ. In breathy voicing, the vocal folds vibrate without much contact, thus the glottis is open for a relatively longer portion of each glottal cycle, resulting in a high OQ. In modal voicing, the vocal folds are open during part of each glottal cycle, resulting in the OQ between those for the creaky voicing and for the breathy voicing.

Other relevant cues for the identification of voice quality, especially creaky voice, include aperiodicity, due to the slow and aperiodic glottal pulses in creaky phonation. A couple of measures can be used to quantify the degree of aperiodicity in the glottal source. One is "jitter", which quantifies the variation in the duration of successive fundamental frequency cycles. Jitter values are higher during creaky phonation than other phonation types. The other is mean autocorrelation ratio. Mean autocorrelation ratio is a temporal measure that quantifies the periodicity of the glottal pulses, which is used in our experiment, as will be detailed in section 3.2.

3. Voice quality decision

3.1 Corpus

Switchboard is a corpus of orthographically transcribed spontaneous telephone conversations between strangers (Godfrey et al. 1992). The corpus is designed mainly to be used in developing robust Automatic Speech Recognition. The corpus consists of more than 300 hours of recorded speech spoken by more than 500 speakers of both genders over the phone. Our analysis is based on a subset of the Switchboard files (12 hours) containing one or more utterance units (10-50 words) from each talker in the corpus. Phone transcriptions are obtained by forced alignment using the word transcription and dictionary. In general, the quality of the recorded speech, which is sampled at 8kHz, is much inferior to speech samples recorded in the phonetics laboratory. Although ITU (International Telecommunication Union) standards only require the telephone network to reproduce speech faithfully between 300Hz and 3500Hz (e.g., ITU Standard 1993), our observations indicate that most signals in Switchboard reproduce harmonics of the fundamental frequency faithfully at frequencies as low as 120Hz. This conclusion is supported by the results of Yoon et al. (2005), who demonstrated that measures of H1-H2 acquired from telephone-band speech are predictive of subjective voice quality measures at a significance level of $p < 0.001$. Post-hoc analysis of Yoon et al.'s results suggests that H1-H2 is an accurate measure of glottalization for female talkers in Switchboard, but is less accurate for male talkers, who often produce speech with F0 < 120Hz. The low quality of telephone-band speech is also known to affect pitch tracking; as noted in Taylor (2000), pitch tracking algorithms known to be reliable for laboratory-recorded speech often fail to extract an F0 during regions perceived as voiced from the Switchboard corpus.

---

proposed by Hanson (1997), and denotes the measure H1-H2 is corrected for the effects of the first formant (F1). See Hanson (1997) and Hanson and Chuang(1999) for the rationale and procedure of obtaining H1*-H2*.

3.2 Feature extraction and voice quality decision

As mentioned above, the Switchboard corpus has the drawback that the recordings are band-limited signals. The voice quality of creakiness is correlated with low F0, which hinders accurate extraction of harmonic structure if the F0 falls below 120Hz. This is because harmonics are any whole-number multiple of F0. To enable a voice quality decision for signals with F0 below 120Hz, we use a combination of two measures: H1-H2 (a spectral measure, occasionally corrupted by the telephone channel) and mean autocorrelation ratio (a temporal measure, relatively uncorrupted by the telephone channel) in the decision algorithm for voice quality.

We use Praat (Boersma and Weenink 2005) to extract the spectral and temporal features that serve as cues to voice quality. First, intensity normalization is applied to each wave file. Following intensity normalization, inverse LPC filtering (Markel 1972) is applied to remove effects of the vocal tract on source spectrum and waveform.

From the intensity-normalized, inverse-filtered signal, minimum F0, mean F0, and maximum F0 are derived over each file. These three values are used to set ceiling and floor thresholds for short-term autocorrelation F0 extraction, and to set a window that is dynamically sized to contain at least four glottal pulses. F0 and mean autocorrelation ratio are calculated on the intensity-normalized, inverse-filtered signal, using the autocorrelation method developed by Boersma (1993). The unbiased autocorrelation function $r_x(\tau)$ of a speech signal $x(t)$ over a window $w(t)$ is defined as in (2):

$$r_x(\tau) \approx \frac{\int x(t)x(t+\tau)dt}{\int w(t)w(t+\tau)dt} \qquad (2)$$

where τ is a time lag. The mean autocorrelation ratio is obtained by the following formula (3):

$$\overline{r}_x = \left\langle \max_{\tau} \frac{r_x(\tau)}{r_x(0)} \right\rangle \qquad (3)$$

where the angle brackets indicate averaging over all windowed segments, which are extracted at a timestep of 10ms. The range of the mean autocorrelation ratio is from 0 to 1, where 1 indicates a perfect match, and 0 indicates no match of the windowed signal and any shifted version. Harmonic structure is determined through spectral analysis using FFT and long term average spectrum (LTAS) analyses applied to the intensity-normalized, inverse filtered signal.

H1 and H2 are estimated by taking the maximum amplitudes of the spectrum within 60 Hz windows centered at F0 and 2×F0, respectively, as in (4):[7]

$$H1-H2 = \max_{-60<\delta_1<60} 20\log_{10}|X(F_0+\delta_1)| - \max_{-60<\delta_2<60} 20\log_{10}|X(2F_0+\delta_2)| \qquad (4)$$

[7] Because the input speech is inverse filtered so that the effect of resonant frequencies are minimized, if not completely eliminated, we didn't apply any correction regarding formants to H1-H2, as is suggested in Hanson (1997) (cf. H1*-H2*)

where $X(f)$ is the FFT spectrum at frequency f.

Yoon et al. (2005) previously used spectral features including H1-H2 to classify subjective voice quality with 75% accuracy. Subjective voice quality labels used in that experiment are not available for the research reported in this paper. In the current work, interactively-determined thresholds are used to divide the two-dimensional feature space [r_x; $H1$-$H2$] into a set of voice-quality-related objective categories, as follows.

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 $H1$-$H2$ < -15dB, or for which $H1$ - $H2$ < 0 and r_x < 0.7. All other frames are labeled with an objective category label called "modal." Figure 5 illustrates an example of objectively labeled creaky voice on the sonorant [er]. The waveform in the top tier is divided into 10ms intervals in the bottom tier. The voiceless, non-creaky, or creaky label is assigned to each 10ms frame based on the above-mentioned criteria. Within each sonorant phone, whose boundaries we obtained through forced alignment, if more frames indicate creaky category than any other category, the phone itself is assigned creaky label ("_cr"). For our experiment, we do not consider the voice quality variation for obstruents such as stops and fricatives, and only sonorants are eligible to be assigned the creaky label.

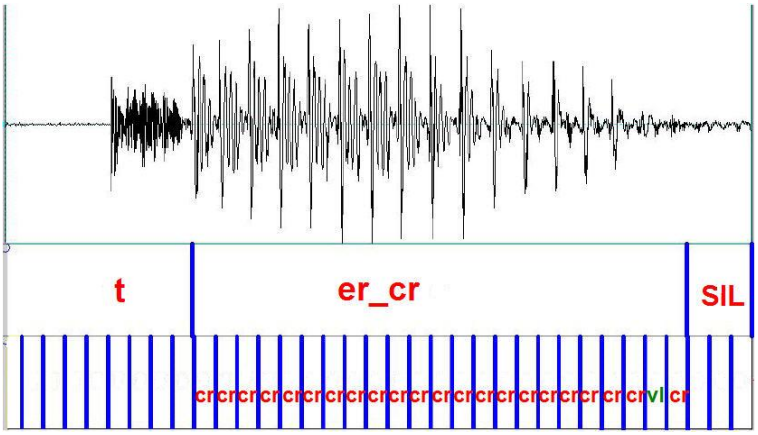


Figure 5: Example of a sonorant /er/with objective creaky label

4. Voice quality distinction reflected in PLP coefficients

As discussed in Section 2, the acoustic measures we extracted (see Section 3) are correlated with the voice quality of creakiness. These features (i.e., H1-H2 and mean autocorrelation ratio) are not a standard input to speech recognition systems. Instead, PLP (Perceptual Linear Predictive) coefficients are usually used as standard input features. There are two ways of incorporating the features related to the voice quality into a speech recognition system: (1) appending the voice quality related features to the standard PLP coefficients, or (2) modeling phones of different voice qualities separately as allophonic variants, while not modifying standard feature vectors. In order to justify the latter approach, it is necessary first to determine whether the voice quality categories are predictive of the standard speech recognition feature vectors such as PLP. This section describes an experiment designed to determine whether or not PLP coefficients are sufficient to distinguish between creaky and non-creaky examples of any given sonorant phone.

The PLP (Perceptual Linear Predictive) cepstrum is an auditory-like cepstrum that combines the frequency-dependent smoothing of MFCC (mel-frequency cepstral coefficients)

with the peak-focused smoothing of LPC (Hermansky 1990). In our work, thirty-nine PLP coefficients are extracted over a window size of 25ms with a timestep of 10ms. PLP coefficients, as shown in the second figure in Figure 6, typically perform well for speech recognition purposes, even with noisy (low SNR) signals. In order to show that the voice quality distinction based on H1-H2 and the mean autocorrelation ratio is also reflected in the acoustic features used in speech recognition, such as PLP coefficients, this section reports the results of a validation test using SVM (Support Vector Machine) classification.

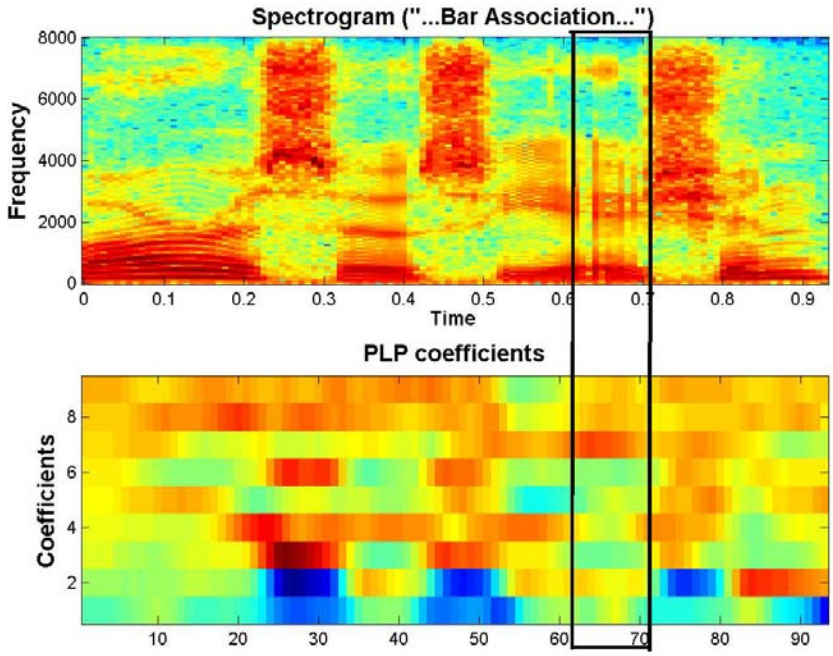


Figure 6: An example of spectrogram and graphical representation of the PLP coefficients. In the spectrogram shown in the first figure, the rectangular region between 0.6 and 0.7 in the x-axis of the upper figure indicates that the speech corresponding to [ei] in the word *Associ**a**tion* is produced with creakiness. This paper investigates whether the creakiness characteristic is reflected in the input feature vectors of PLP coefficients, which is graphically represented in the second figure.

SVM is a machine learning algorithm that seeks to find the optimal mapping function $y = f(x, \alpha)$, where y is an output category (e.g., either modal or creaky phones), x is an input feature vector (e.g., PLP coefficients), and α is a set of adjustable model parameters. The optimality is defined by minimizing the structural error of the classification. We use SVM with a non-linear kernel because we assume that the category boundary between model and creaky phones is nonlinear in the feature space of PLP coefficients.

We conduct an experiment to classify non-creaky phonation versus creaky phonation for each sonorant (i.e., vowel, semi-vowel, nasal or lateral). The phone-aligned transcription for each file is obtained using HTK (Young et al. 2005), and aligned against the voice quality label sequences given by the frame-level voice quality decisions described before. For each sonorant segment, if more frames indicate creakiness than the other voice qualities (i.e., modal or voiceless), the phone is labeled as creaky. We divide the 12 hour Switchboard subset into a

training candidate pool (90%) and a testing candidate pool (10%). Then for each sonorant phone from the training candidate pool, we extract a subset of the non-creaky tokens that is equal in size to the creaky tokens for the same phone, based on the creakiness label resulting from the decision scheme. These non-creaky and creaky tokens compose the training data for each sonorant. The testing data for each sonorant are similarly generated from the testing candidate pool, which also have equal numbers of creaky and non-creaky tokens and no overlap with the training data. We use the SVM toolkit LibSVM (Chang and Lin 2004) to train separate binary classifiers for each sonorant; each classifier distinguishes between creaky and non-creaky examples of the phone. Classifiers are tested using the testing data, for each sonorant separately. The classification accuracies obtained from the testing data for each sonorant are reported in Table 1.

Table 1: SVM classification of voice qualities for each phone: The first and third columns list the creaky (indicated by cr) versus non-creaky phone labels, in ARPABET notation. The second and fourth columns list the accuracy of a classifier trained to distinguish between creaky and non-creaky examples of the specified phone.

Phones		**Accuracy**	**Phones**		**Accuracy**
uh	uh_cr	74.47%	w	w cr	69.91%
dr	er_cr	73.26%	ih	ih cr	69.75%
aw	aw cr	73.26%	ow	ow cr	69.09%
eh	eh cr	71.93%	y	y cr	68.45 %
ae	ae cr	71.52%	l	l_cr	68.23 %
uw	uw_cr	71.42%	ao	ao_cr	68.04 %
iy	iy_cr	70.51%	m	m_cr	67.79 %
ey	ey_cr	70.50 %	ax	ax_cr	67.24 %
ay	ay_cr	70.37 %	el	el_cr	66.85 %
ah	ah_cr	70.14 %	r	r_cr	66.36 %
aa	aa_cr	70.13 %	oy	oy_cr	63.24 %
ng	ng_cr	70.05 %	en	en_cr	58.19 %
n	n_cr	70.03 %			

As shown in Table 1, the PLP coefficients are correctly classified with an overall accuracy of 58% to 74% (with an average overall accuracy of 69.23%). Chance performance is 50%. An average of 19.23% improvement, relative to chance, suggests that the voice quality decision is reflected to some degree in the PLP coefficients. Based on this finding, we conclude that it should be possible to design a speech recognition system that distinguishes between creaky and non-creaky examples of each sonorant phone using only PLP coefficients as an acoustic observation.

5. Voice quality dependent speech recognition

The goal of a speech recognition system is to find the word sequence that maximizes the posterior probability of the word sequence $\mathbf{W} = (w_1, w_2, \cdots, w_M)$, given the observations $\mathbf{O} = (o_1, o_2, \cdots, o_T)$:

$$\hat{W} = \arg\max_{\mathbf{w}} p(\mathbf{W} \mid \mathbf{O}) \tag{5}$$

Using Bayes rule and the fact that $p(\mathbf{O})$ is not affected by $\mathbf{W}$,

$$\begin{aligned} \hat{W} &= \arg\max_{\mathbf{w}} \frac{p(\mathbf{O} \mid \mathbf{W}) p(\mathbf{W})}{p(\mathbf{O})} \\ &= \arg\max_{\mathbf{w}} p(\mathbf{O} \mid \mathbf{W}) p(\mathbf{W}) \end{aligned} \tag{6}$$

Sub-word units $\mathbf{Q} = (q_1, q_2, \cdots, q_L)$, such as phones, are usually essential to large vocabulary speech recognition, therefore we can rewrite formula (7) as:

$$\begin{aligned} \hat{W} &= \arg\max_{\mathbf{w}} p(\mathbf{O} \mid \mathbf{W}) p(\mathbf{W}) \\ &\approx \arg\max_{\mathbf{w}} [\max_{Q} p(\mathbf{O} \mid \mathbf{Q}) p(\mathbf{Q} \mid \mathbf{W}) p(\mathbf{W})] \end{aligned} \tag{7}$$

The general automatic speech recognition architecture is shown in Figure 7. The post probability of each word sequence hypothesis $\mathbf{W}$ is calculated according to three components: the acoustic model $p(\mathbf{O} \mid \mathbf{Q})$, the pronunciation model $p(\mathbf{Q} \mid \mathbf{W})$ and the language model $p(\mathbf{W})$.

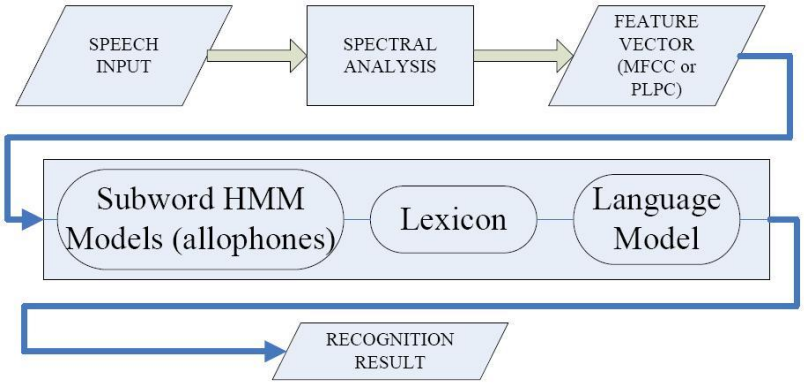


Figure 7: General automatic speech recognition architecture

In a typical speech recognition system, the observation vectors $\bar{O}$ are PLP (Perceptual Linear Predictive) coefficients or MFCC (Mel Frequency Cepstral Coefficients), plus their energy, all computed over a window size of 25ms at a time step of 10ms, and their first order and second order regression coefficients, referred to as delta and delta-delta (or acceleration) coefficients.

The acoustic model $P(\mathbf{O} \mid \mathbf{Q})$ is usually a set of left-to-right hidden Markov models (HMMs), each modeling the acoustics of a particular sub-word unit such as a phone, as in Figure 8:

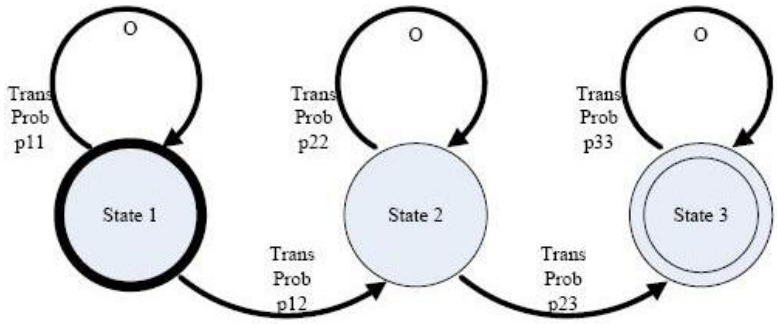


Figure 8: Left-to-right hidden Markov model

In a left-to-right HMM, state transitions occur from a state either to itself or to the following state. These state transition probabilities describe, from a probabilistic point of view, how long each part of the sub-word unit *q* should be. For each of the states, there is one Gaussian-mixture distribution describing the state-conditioned observation distributions.

The pronunciation model $p(\mathbf{Q} \mid \mathbf{W})$ typically maps a word to either phones or triphones (allophones in particular contexts). In this paper, we are using a deterministic pronunciation model, i.e. mapping each word to a fixed sequence of triphones.

The language model $p(\mathbf{W})$ is usually the *n*-gram model: the probability of a particular word in the word sequence is conditioned on the previous *n-1* words.

$$p(w_1 w_2 \cdots w_m) = p(w_1) \cdots p(w_{n-1}) \prod_{i=n}^{m} p(w_i \mid w_{i-n+1} \cdots w_{i-1}) \tag{8}$$

For example, the simple bigram language model is as follows:

$$p(w_1 w_2 \cdots w_m) = p(w_1) \prod_{i=2}^{m} p(w_i \mid w_{i-1}) \tag{9}$$

5.1 Baseline system

We build a triphone-clustered HMM-based speech recognition system as the baseline system using HTK (Young et al. 2005). This system uses a deterministic pronunciation model, also called dictionary, and a bigram language model, but a sophisticated acoustic model, which will be detailed in the following paragraphs.

Every phone is represented by a large number of partially independent triphone models. All triphones that represent the same base phoneme use the same transition probability matrix; we say that their transition probability matrices are “tied.” The observation probability density functions associated with the first, second, or third state of any given pair of triphones may also be tied together, as shown in Figure 9.

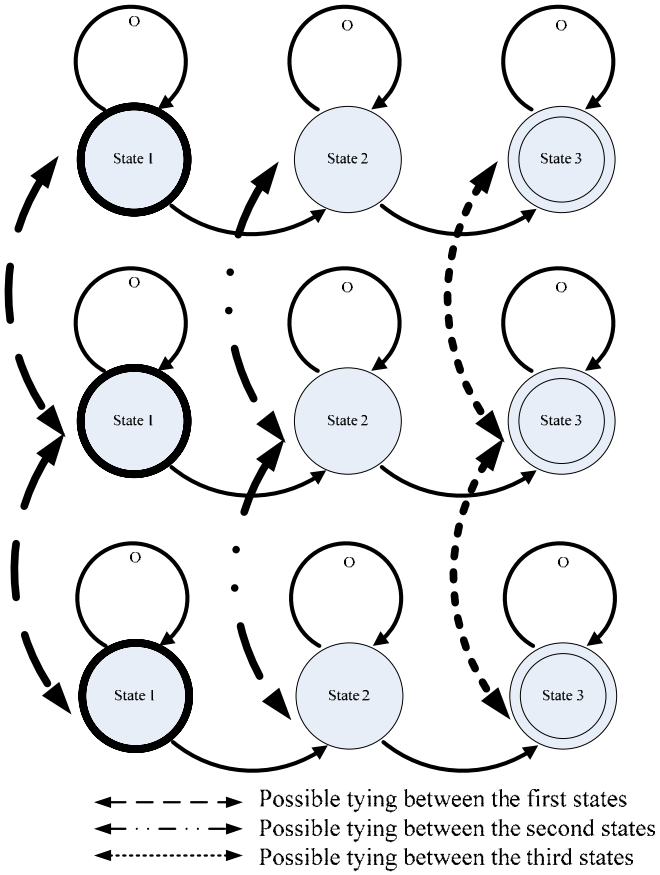


Figure 9: Tying options of counterpart states in HMMs representing allophones of the same base phoneme.

Allophones of the same base phoneme are tied together in allophone sets. Each allophone set corresponds to one of the leaves in a binary tree. The phonetic binary clustering tree (Figure 10) begins with a root node comprising all allophones of a given base phoneme label. At each level of the tree, the allophones belonging to the next higher level are split into two categories based on a question about the phonological features of the left context phone or the right context phone. The tree is grown from root to leaf (or from top down in Figure 10), with all corresponding states of allophones placed at the root node initially. At each non-leaf node, the splitting question is selected from a pool of binary questions to maximize the increase in the likelihood of training data given the model. In this way, phonetic contexts that induce the most allophonic variation are placed nearest to the tree root. Once the maximum likelihood increase at a particular node is smaller than a threshold, this node will not be further split and all states in that node will be tied together.

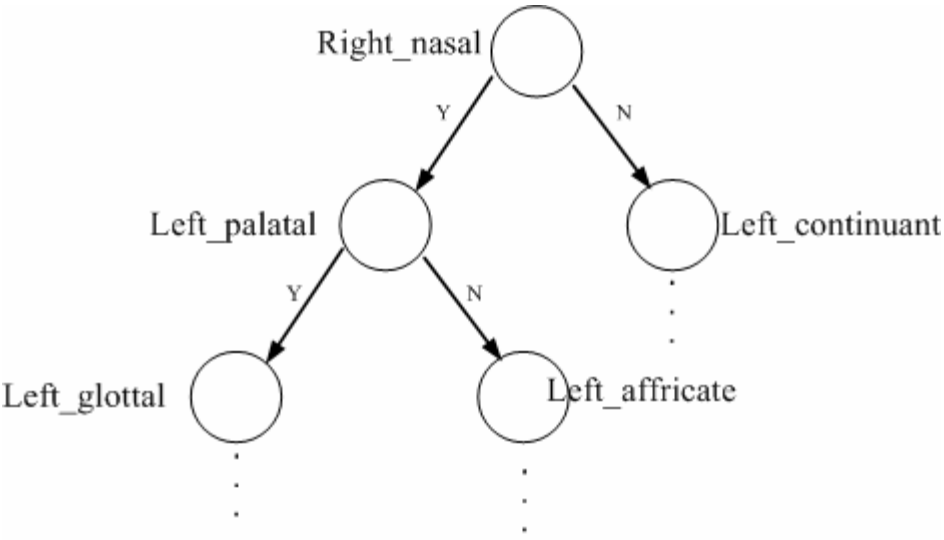


Figure 10: Binary clustering tree (an example of the near-root part of the binary clustering tree for the third emitting state of vowel /ae/)

It is necessary to deal with triphones unseen in the training data but maybe existing in testing data. These triphone models are synthesized, after the model is fully trained, by tying the states of the HMM to three particular states from seen allophones, chosen according to the unseen triphone's answer to binary questions in the clustering tree. In other words, a synthesized state is tied to all the states in a particular leaf node of the clustering tree.

After state tying is completed, the number of Gaussians in each mixture Gaussian observation distribution is repeatedly incremented, with further mean and variance estimation following each increment, thus achieving observation distributions that better reflect the characteristics of the allophones.

5.2 VQ-ASR system

The Voice Quality Automatic Speech Recognition (VQ-ASR) system incorporates into the baseline system binary voice quality information (creaky or non-creaky) for every sonorant phone.

Inclusion of Voice Quality Information: We use forced alignment to obtain phone boundaries for the phonemes specified in the canonical dictionary entry for each word listed in the Switchboard word transcription. This phone-aligned transcription is aligned against the voice quality label sequences given by the frame-level voice quality decisions described in subsection 3.2. For a vowel, semi-vowel or nasal, if more frames indicate creakiness than the other voice qualities (i.e., modal or voiceless), a "creakiness label" is attached to this phonation (See Figure 5).

Given these creakiness-labeled phone transcriptions and corresponding wave files, we use the Baum-Welch algorithm to do an embedded estimation of all the allophone HMMs involved in these transcriptions. For every training utterance, the HMMs corresponding to phones present in that utterance are concatenated according to the transcription, and estimated together instead of separately. Thus, we can get one HMM for each allophone, defined on its own phone identity and its context, both in terms of phonetics and voice quality. The creakiness of a phone is modeled as part of the phone's context, rather than being part of the base phoneme label,

thus creaky and non-creaky versions of the same phoneme are eligible to be clustered together by the triphone clustering algorithm exemplified in Figure 10. Figure 11 illustrates how voice quality knowledge is incorporated in the training transcription.

Word transcription:	SAY	YOU	DID
Phone transcription:	s ey sp	y uw sp	d ih d sp
Creakiness labels:		cr	cr
Creakiness-labeled phone transcription:	s ey sp	y uw_cr sp	d ih_cr d sp
Allophone transcription	s+ey s-ey sp y+uw_cr y-uw_cr sp d+ih_cr d-ih_cr+d ih_cr-d sp		
Allophone Transcription (phonetic / voice quality context)	s+ey s-ey sp y+uw_cr y_cr-uw sp d+ih_cr d_cr-ih+d_cr ih_cr-d sp		

Figure 11: Conversion from the word transcription to the transcription of allophones defined on phone identify and phonetic/voice quality context. ("_cr" represents the "creakiness label".)

Recognition Dictionary with Voice Quality Information: To perform speech recognition using voice quality information, we need to map the voice quality dependent allophone sequences to word sequences. While we wish to take advantage of explicit acoustic modeling of voice quality variation, such variation does not impact word identity (in English). Therefore, we need a new dictionary containing all possible pronunciations of the same word, with all of the different possible voice quality settings. For example, for "bat b+ae b-ae+t ae-t" in the baseline system dictionary, as in Figure 12(a), the dictionary in a VQ-ASR system should have two entries "bat b+ae b-ae+t ae-t" and "bat b+ae_cr b_cr-ae+t_cr ae_cr-t", as in Figure 12(c).

Word:	bat		
Phones:	b	ae	t
(a) Triphones:	b+ae	b-ae+t	ae-t
(b) Triphones:	b+ae	b-ae+t	ae-t
(with VQ Info)	b+ae_cr	b-ae_cr+t	ae_cr-t
(c) Triphones:	b+ae	b-ae+t	ae-t
(VQ context)	b+ae_cr	b_cr-ae+t_cr	ae_cr-t

Figure 12: Recognition dictionary with voice quality information (example: the word "bat")

Reduction of the Number of Parameters: The number of triphones increases dramatically, as the creakiness label can be attached to one or both of the neighboring phones for each triphone. To reduce the number of parameters, we include allophones with different phonetic/voice quality context in the same binary decision tree in the triphone clustering process (Figure 13). By tying transition matrices of all allophones, tying states of some allophones using a tree-based clustering technique, and synthesizing unseen triphones in the same way as the baseline system, we build the VQ-ASR system with an almost identical number of parameters to that in the baseline system, despite the increase in the number of triphones. This is necessary, because any increase in the number of model parameters will have a tendency to improve recognition performance, which would make the comparison between the VQ-ASR system and the baseline system less accurate.

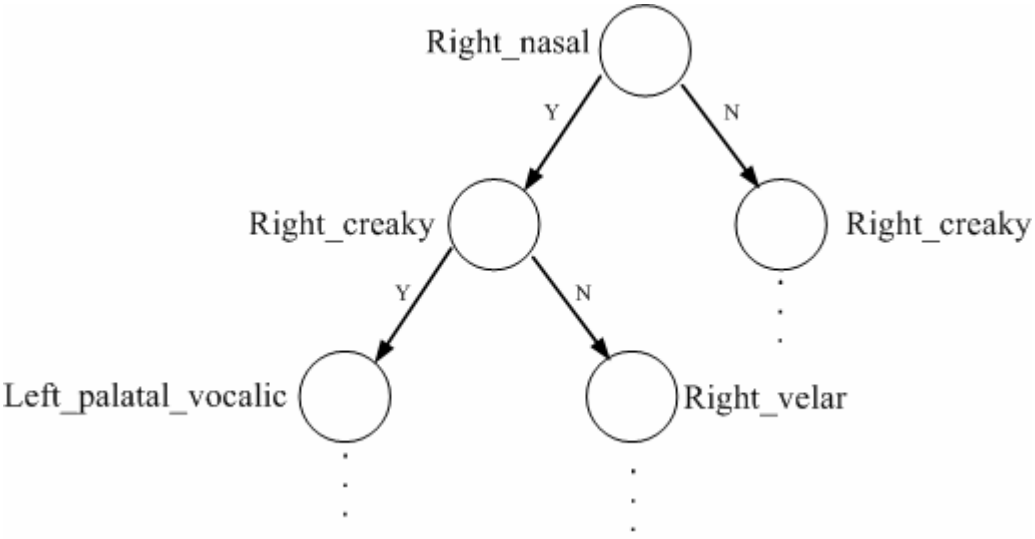


Figure 13: Binary clustering tree showing the effect of creakiness. (an example of the near root part of the binary clustering tree for the third emitting state of vowel /ae/, showing that creakiness context is more salient than most phonetic context)

6 Experimental results

Word recognition accuracies of the voice quality dependent and voice quality independent speech recognition systems are shown in Table (2). In our experiment, both systems are prototype ASR systems, trained and tested on the 12 hour subset of Switchboard[8]. The comparison of the results in Table (2) is made under the condition of (i) tied transition probabilities for all allophones and (ii) an almost identical number of states for both systems. This allows for a stringent comparison between systems with a nearly equal number of parameters.

[8] The two systems are designed to identify the impact of voice quality dependence, therefore not comparable to systems trained on much larger amounts of data (e.g., Luo and Jelinek 1999; Sundaram et al. 2000).

Table 2: Word recognition accuracy for the voice quality dependent and voice quality independent recognizers. The number of Gaussians in each Gaussian mixture is given in the first column. %Correctness is equal to the percentage of the reference labels that were correctly recognized. %Accuracy is a more comprehensive measure of recognizer quality that penalizes insertion errors.

Mixture	Baseline		VQ-ASR	
	% Correctness	%Accuracy	% Correctness	%Accuracy
3	45.81	39.28	46.42	39.35
9	52.77	45.31	52.77	46.01
19	52.88	46.82	55.41	48.63

Two evaluation metrics are used: %Correctness and %Accuracy, defined as

$$\%\text{Correctness}=\frac{N-D-S}{N}\times 100$$
$$\%\text{Accuracy}=\frac{N-D-S-I}{N}\times 100 \qquad (10)$$

where N is the number of tokens (i.e. words) in the reference transcriptions that have been reserved as a test dataset for the evaluation purpose, D, the number of deletion errors, S, the number of substitution errors, and I, the number of insertion errors. The %Correctness penalizes deletion errors and substitution errors deviating from the reference transcriptions; %Accuracy also penalizes insertion errors. Word error rate (WER), another widely used evaluation metric, is equal to 100-%Accuracy.

As seen in Table 2, when voice quality information is incorporated in the speech recognition system, the percentage of words correctly recognized by the system increases by approximately 0.86% on average and the word accuracy increases by approximately 1.05% on average. It is worth noting that as the number of Gaussians per mixture increases to 19, the improvement in the percentage of words correctly recognized increases to 2.53%, and the improvement in the word accuracy increases to 1.81%.

7 Discussion and conclusion

In this paper, we have shown that a voice quality decision based on H1-H2 as a measure of harmonic structure, and the mean autocorrelation ratio as a measure of temporal periodicity, provides useful allophonic information to an automatic speech recognizer. Such voice quality information can be effectively incorporated into an HMM-based automatic speech recognition system, resulting in improved word recognition accuracy.

As the number of Gaussian components per state of the HMM increases, the VQ-ASR system surpasses the baseline system by an increasingly greater extent. Given that the number of untied states and the number of transition probabilities in the HMMs in both systems are identical, it follows that the VQ-ASR system benefits more from an increasingly precise observation PDF (probability density function), compared to the baseline system. Although we

don't know why added mixtures might help the VQ-ASR more than the baseline, we speculate that there must be an interaction between the phonetic information provided by voice quality labels, and the phonetic information provided by triphone context. Perhaps the acoustic region represented by each VQ-ASR allophone is fully mapped out by a precise observation PDF to an extent not possible with standard triphones.

Similar word recognition accuracy improvements have been shown for allophone models dependent on prosodic context (Borys 2003). Glottalization has been shown to be correlated with prosodic context (e.g., Redi and Shattuck-Hufnagel 2001), thus there is reason to believe that an ASR trained to be sensitive to both glottalization and prosodic context may have super-additive word recognition accuracy improvements.

Acknowledgment

This work is supported by NSF (IIS-0414117). Statements in this paper reflect the opinions and conclusions of the authors, and are not necessarily endorsed by the NSF.

References

BICKLEY, C. 1982. Acoustic analysis and perception of breathy vowels. *Speech Communication Group Working Papers*, Research Laboratory of Electronics, MIT, 73-93.

BOERSMA, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampling sound. In: *Proceedings of the Institute of Phonetic Sciences*. No. 17. University of Amsterdam.

BOERSMA, P., WEENINK, D. 2005. *Praat: doing phonetics by computer* (Version 4.3.19). [computer program], http://www.praat.org.

BORYS, S. 2003. The importance of prosodic factors in phoneme modeling with applications to speech recognition. In: *HLT/NAACL student session*. Edmonton.

CHANG, C.-C., LIN, C.-J. 2004. *Libsvm: a library for support vector machine*. System documentation, <http://www.csie.ntu.edu.tw/ ˜cjlin/libsvm>.

DILLEY, L., SHATTUCK-HUFNAGEL, S., OSTENDORF, M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24: 423–444.

EPSTEIN, M. A. 2002. *Voice Quality and Prosody in English*. Ph.D. dissertation, UCLA, California, LA.

FANT, G. 1960. *Acoustic Theory of speech production*. The Hague: Mouton

FANT, G. 1999. The voice source in connected speech. *Speech Communication* 22: 125-139.

FANT, G., KRUCKENBERG, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2: 1-83, Speech, Music and Hearing, Royal Institute of Technology, Stockholm,

FISCHER-JØRGENSEN, E. 1967. Phonetic analysis of breathy (murmured) vowels. *Indian Linguistics* 28: 71-139.

GERRATT, B., KREIMAN, J. 2001. Towards a taxonomy of nonmodal phonation. *Journal of Phonetics* 29: 365–381.

GOBL, C. 2003. *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.

GODFREY, J., HOLLIMAN, E., MCDANIEL, J. 1992. Telephone speech corpus for research and development. In: *Proceedings of ICASSP*. San Francisco, CA.

GORDON, M. 1996. The phonetic structures of Hupa, *UCLA Working Papers in Phonetics* 93: 164-187.

GORDON, M., LADEFOGED, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29: 383–406.

HANSON, H. 1997. Glottal characteristics of female speakers: acoustic correlates. *Journal of the Acoustical Society of America* 101: 466-481.

HANSON, H., CHUANG, E. 1999. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America* 106: 1697–1714.

HANSON, H., STEVENS, K.N., KUO, J., CHEN, M., & SLIFKA, J. 2001. Towards models of phonation. *Journal of Phonetics* 29: 451-480.

HERMANSKY, H. 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America* 87: 1738–1752.

HUFFMAN, M.K. 1987. Measures of phonation type in Hmong. *Journal of the Acoustical Society of America* 81: 495-504.

INTERNATIONAL TELECOMMUNICATION UNION (ITU) STANDARD G.711, 1993. Pulse code modulation (pcm) of voice frequencies.

JIANFEN, C., MADDIESON, I. 1989. An exploration of phonation types in Wu dialects of Chinese. *UCLA Working Papers in Phonetics* 72: 139-160.

KUSHAN, S., SLIFKA, J. 2006. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In: *ICSA International Conference on Speech Prosody*. Dresden, Germany.

LAVER, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge, Cambridge University Press.

LUO, X., JELINEK, F. 1999. Probabilistic classification of hmm states for large vocabulary continuous speech recognition. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 353–356.

MADDIESON, I., HESS, S. 1987. The effects of F0 of the linguistic use of phonation type. *UCLA Working Papers in Phonetics* 67: 112–118.

MARKEL, J. D. 1972. The sift algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics* 20: 367–377.

NÍ CHASAIDE, A., GOBL, C. 1997. Voice source variation. In: Hardcastle,W., Laver, J. (Eds.), *The Handbook of Phonetic Sciences*. Blackwell Publishers, Oxford, pp. 1–11.

PIERREHUMBERT, J. 1989. A preliminary study of the consequences of intonation for the voice source. *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm 4: 23-36.

REDI, L., SHATTUCK-HUFNAGEL, S. 2001. Variation in the rate of glottalization in normal speakers. *Journal of Phonetics* 29: 407–427.

SUNDARAM, R., GANAPATHIRAJU, A., HAMAKER, J., PICONE, J. 2000. ISIP 2000 conversational speech evaluation system. In: *NIST Evaluation of Conversational Speech Recognition over the Telephone*.

SWERT, M.; VELDHUIS, R. 2001. The effect of speech melody on voice quality. *Speech Communication* 33: 297-303.

TAYLOR, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the*

Acoustical Society of America 107: 1697–1714.
THONGKUM, T.L. 1987. Another look at the register distinction in Mon. *UCLA Working Papers in Phonetics* 67: 132-165.
YOON, T.-J., COLE, J., HASEGAWA-JOHNSON, M., SHIH, C. 2005. Acoustic correlates of nonmodal phonation in telephone speech. *Journal of the Acoustical Society of America* 117: 2621.
YOUNG, S., EVERMANN, G., GALES, M., HAIN, T., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V., WOODLAND, P. 2005. *The HTK Book* (version 3.3). Technical Report, Cambridge University Engineering Department, Cambridge, UK.

Tae-Jin Yoon
Department of Linguistics, 4080 FLB
University of Illinois at Urbana-Champaign
707 S. Mathews Ave.
Urbana, IL 61801 U.S.A.
tyoon@uiuc.edu

Xiaodan Zhuang
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
405 N. Mathews Ave.
Urbana, IL 61801 U.S.A.
xzhuang2@uiuc.edu

Jennifer Cole
Department of Linguistics, 4080 FLB
University of Illinois at Urbana-Champaign
707 S. Mathews Ave.
Urbana, IL 61801 U.S.A.
jscole@uiuc.edu

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
405 N. Mathews Ave.
Urbana, IL 61801 U.S.A.
jhasegaw@uiuc.edu