Voice Quality Dependent Speech Recognition

Tae-Jin Yoon[1], Xiaodan Zhuang[2], Jennifer Cole[1], and Mark Hasegawa-Johnson[2]
Department of Linguistics[1]; Department of Electrical and Computer Engineering[2]
University of Illinois at Urbana-Champaign
{tyoon; xzhuang2; jscole; jhasegaw}@uiuc.edu

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 harmonic structure (H1-H2) and the mean autocorrelation ratio of each phone. Results from a Support Vector Machine (SVM) classification experiment show that these features are predictive of Perceptual Linear Predictive Cepstra (PLPC) used in speech recognition. We further demonstrate that by incorporating voice quality knowledge into a speech recognition system, we can improve word recognition accuracy.

1 Introduction

Through modulation in source and filter characteristics, speech conveys both linguistic and paralinguistic information. Fundamental frequency (F_0) and harmonic structure are important factors in encoding lexical contrast and allophonic variation related to laryngeal features. It has been widely noted that there is a relationship between F_0 and voice quality. However, F_0 is not always a strong indicator of voice quality, as shown by studies of English that fail to show a strong correlation between any glottal parameters and F_0 (Epstein, 2002). On the other hand, information obtained from spectral structure has been shown to be more reliable for the discrimination of non-modal from modal phonation (Hanson, 1997; Epstein, 2002) .

In this paper, we address the viability of voice quality analysis for large corpora of low quality recorded speech by labeling the voice quality using both harmonic structure (a spectral measure, occasionally corrupted by the telephone channel) and mean autocorrelation ratio (a temporal measure, relatively uncorrupted by the telephone channel). A validation test using Support Vector Machines (SVM) demonstrates that these voice quality measures are correlated with the average PLP (Perceptual Linear Predictive) cepstrum of a phone. We show that an automatic speech recognizer that incorporates voice quality information into the system performs better than a complexity-matched baseline system that does not consider the voice quality distinction.

2 Voice Quality Decision

Switchboard is a corpus of orthographically transcribed spontaneous telephone conversations between strangers. We use a subset of the Switchboard files (12 hours) containing one or more utterance units (10-50 words) from each talker in the corpus. The Switchboard corpus has the drawback that the recordings are bandlimited signals (120Hz-4kHz). The voice quality of creakiness is correlated with low F_0, which hinders accurate extraction of harmonic structure if the F_0 falls below 120Hz. To enable a voice quality decision for signals with F_0 below 120Hz, we use a combination of two measures: H1-H2 and mean autocorrelation ratio in the decision algorithm for voice quality (Boersma and Weenink, 2005; Hanson, 1997; Boersma, 1993). Interactively-determined thresholds are used to divide the two-dimensional feature space $[\bar{r}_x, H1 - H2]$ 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 phonation” category includes all frames for which $H1 - H2 < -15$dB, or for which $H1 - H2 < 0$ and $\bar{r}_x < 0.7$. All other frames are labeled with an objective category label called “non-creaky phonation.”

3 Relationship with PLPC

PLPC (Hermansky, 1990) is an auditory-like spectrum that combines together the frequency-dependent smoothing of MFSC (mel-frequency spectral coefficients) with the peak-focused smoothing of LPC. In order to show that the voice quality distinction based on H1-H2 and the mean autocorrelation ratio is also reflected in the PLPC used in speech recognition, we conducted an experiment to classify non-creaky phonation versus creaky phonation for each sonorant (i.e., vowel, semi-vowel, nasal or lateral) using SVM (Chang and Lin, 2005). The classification accuracies obtained from the testing data for each sonorant are reported in Table (1). Our purpose here is to verify whether there are acoustic differences in the PLPC coefficients that reflect the voice quality distinction we identify using the knowledge-based method described in the previous section. We do not attempt to optimize the SVM classification of creaky versus non-creaky phones in this experiment. Therefore, the default parameter setting of the radial basis function (RBF) in LibSVM is used without modification.

An average of 19.23 % of improvement is achieved in the experiment. The result suggests that the voice quality decision is reliably reflected in the PLPC features, on which basis we conducted a speech recognition experiment based on the PLPC correlates of voice quality information.

4 VQ-ASR

We build a triphone-clustered HMM-based speech recognition system as the baseline system. The Voice Quality Automatic Speech Recognition (VQ-ASR) system incorporates into the baseline system binary voice quality information (creaky/non-creaky) for every sonorant phone. We use HTK (Young et al., 2005) to obtain the phone-aligned transcription. This phone-aligned transcription is aligned against the voice quality label sequences given by the frame voice quality decisions taken at described before. To perform speech recognition using voice quality information, we need a new dictionary having all possible pronunciations of the same word, with different voice quality settings. We treat the triphones with different voice quality setting as allophones of the same root monophone. By tying transition matrices of all allophones, tying states of some allophones with the help of 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 as the baseline system, despite the increase of triphones. This is necessary, because any increase in model parameters will have a tendency to improve recognition performance, which would make the comparison between the VQ-ASR system and the baseline system inaccurate.

Table 1: *SVM-based voice quality classification for each phone. The first and third columns list the creaky (indicated by _cr) versus non-creaky phones. The second and fourth columns are the overall accuracy of the classification result.*

phones		Accuracy	phones		Accuracy
uh	uh_cr	74.47 %	w	w_cr	69.91 %
er	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 %			

Word recognition accuracies of the voice quality dependent and voice quality independent speech recognition systems are shown in Table (2). As seen in the table, 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 mixtures 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%.

Table 2: *Word recognition accuracy for the voice quality dependent and voice quality independent recognizers. In the first column is the number of mixture components.*

Mixture	Baseline		VQ-Dependent	
	% Correct	Acc.	% Correct	Acc.
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

5 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 incorporated into an HMM-based automatic speech recognition system effectively, resulting in improved word recognition accuracy. As the number of mixture components 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 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.

As the number of mixture components 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 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, such that the triphone clusters generated by the VQ-ASR system cover a compact but not necessarily convex region of acoustic space. If this speculative explanation is correct, then perhaps the compact 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.

6 Acknowledgments

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

References

Paul Boersma 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampling sound. *IFA Proceedings*, 17: 97-100, University of Amsterdam.

Paul Boersma and David Weenink. 2005. *Praat: doing phonetics by computer*. [computer program] `http://www.praat.org`.

C.-C. Chang and C.-J. Lin 2005 *LIBSVM: a library for support vector machines*. `http://www.csie.ntu.edu.tw/˜cjlin/libsvm`.

Melissa Epstein 2002. *Voice Quality and Prosody in English*. Ph.D. dissertation, UCLA.

Helen M. Hanson. 1997. Glottal characteristics of female speakers: acoustic correlates. *J.Acoust.Soc.Am*, 101: 466-481.

Hynek Hermanksy. 1990. Perceptual linear predictive (PLP) analysis of speech *J.Acoust.Soc.Am.*, 87: 1738-1752.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., 2005. *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, Cambridge, UK. `http://htk.eng.cam.uk/`