# Voice Quality Dependent Speech Recognition

Tae-Jin Yoon[†], Xiaodan Zhuang[‡],
Jennifer Cole[†] & Mark Hasegawa-Johnson[‡]

Department of Linguistics[†];
Department of Electrical and Computer Engineering[‡]
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA

tyoon@uiuc.edu, xzhuang2@uiuc.edu,
jscole@uiuc.edu, 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.

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

# 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 (Gordon and Ladefoged, 2001; Maddieson and Hess, 1987). They also play an important role in the expression of prosodic features of stress and intonation (Epstein, 2002; Redi and Shattuck-Hufnagel, 2001). In addition, shifts in $F_0$ and voice quality can signal emotional state or affect, as for example creaky voice is likely to signal the expression of boredom, and breathy voice tends to signal intimacy.

It has been widely noted that there is a relationship between $F_0$ and voice quality. For example, Maddieson and Hess (1987) observe significantly higher $F_0$ for tense vowels in languages that distinguish three phonation types with varying voice quality (Jingpho, Lahu and Yi). However, $F_0$ is not always a reliable 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. 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 $F_0$. They report that spectral slope is the most important feature for discrimination among different phonation types. Ní Chasaide and Gobl (1997) also characterize creaky phonation as having low $F_0$ and irregular glottal pulses. 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), and ii) H2 (i.e., amplitude of the second harmonic) higher than H1.

Among numerous categories of voice quality (e.g., see Gerratt and Kreiman (2001)) , it is known that creakiness (or glottalization) is highly correlated with linguistic structure such that creakiness is more likely to be observed at word, syntactic, or prosodic boundaries (Kushan and Slifka, 2006; Dilley et al., 1996; Redi and Shattuck-Hufnagel, 2001; Epstein, 2002). Since creaky voice can cue word-level and higher juncture, direct modeling of voice quality such as creakiness in speech recognition systems is expected to result in improved word recognition accuracy. And yet, the established importance of spectral structure, and in particular the relative amplitude of the lower harmonics, for voice quality identification calls to question the viability of voice quality analysis for large speech

corpora, especially corpora consisting of low quality recorded speech, such as telephone speech. We address this challenge in the present study by labeling the voice quality of spontaneous telephone speech 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-related measures are correlated with the average PLP cepstrum of phones. We show that a PLC-cepstrum-based 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.

The paper is organized as follows. Section 2 introduces our method of voice quality decision on the Switchboard corpus of telephone conversation speech. Section 3 reports a classification result that shows the voice quality distinctions are reflected in PLPC. Section 4 presents an HMM-based Automatic Speech Recognition System (ASR) that incorporates voice quality knowledge. Section 5 compares the performance of the voice quality dependent recognizer against a baseline system that doesn't distinguish different voice qualities. Section 6 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 decision

## 2.1 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. 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 frequen-

cies 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 $F_0 < 120$Hz. 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 $F_0$ during regions perceived as voiced from the Switchboard corpus.

## 2.2 Feature extraction and voice quality decision

As mentioned above, the Switchboard corpus has the drawback that the recordings are bandlimited signals. 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 (a spectral measure) and mean autocorrelation ratio (a temporal measure) 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 $F_0$, mean $F_0$, and maximum $F_0$ are derived over each file. These three values are used to set ceiling and floor thresholds for short-term autocorrelation $F_0$ extraction, and to set a window that is dynamically sized to contain at least four glottal pulses. $F_0$ 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 (1):

$$r_x(\tau) \triangleq \frac{\int x(t)x(t+\tau)dt}{\int w(t)w(t+\tau)dt} \tag{1}$$

where $\tau$ is a time lag. The mean autocorrelation ratio is obtained by the following

formula (2):

$$\bar{r}_x = \left\langle \max_{\tau} \frac{r_x(\tau)}{r_x(0)} \right\rangle \tag{2}$$

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 $F_0$ and $2\times F_0$, respectively, as in (3):

$$\begin{aligned} H1 - H2 \;\; = \;\; & \max_{-60<\delta_1<60} 20\log_{10}|X(F_0+\delta_1)| \\ & - \max_{-60<\delta_2<60} 20\log_{10}|X(2F_0+\delta_2)| \end{aligned} \tag{3}$$

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

H1 and H2 are related to the Open Quotient (OQ) (Hanson and Chuang, 1999). OQ is defined as the ratio of the time in which the vocal folds are open to the total length of the glottal cycle. In creaky voicing, the vocal folds are held tightly together (though often with low internal tension), resulting in a low 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.

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 $[\bar{r}_x, H1-H2]$ 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 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 Voice quality distinction reflected in PLPC

As discussed in section 1, the acoustic measures we extracted (see section 2) 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, PLPC (Perceptual Linear Predictive Cepstra) or MFCC (Mel Frequency Cepstral 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, as described in section 2, to the standard PLPC or MFCC feature vectors, or (2) modeling phones of different voice qualities separately as allophonic variants, while not modifying standard feature vectors. In the latter approach, which we use in our current experiment, it is necessary to test whether the voice quality related features are related to the standard speech recognition feature vectors.

PLPC (Perceptual Linear Predictive Cepstra) 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 (Hermansky, 1990). In our work, thirty-nine PLPC coefficients are extracted over a window size of 25ms with a timestep of 10ms. PLPC features 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 PLPC, this section reports the results of a validation test using SVM (Support Vector Machine) classification.

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), which implements a statistical learning technique for pattern classification, to perform supervised training of binary classifiers of creaky versus non-creaky phones for each sonorant, and tested the classification over the testing data, for each sonorant separately.

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.

Table 1: *SVM classification of voice qualities 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 results.*

| 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 % | | | |

As shown in the Table (1), the PLPC features are correctly classified with the overall accuracy of 58% to 74% (with an average overall accuracy of 69.23%). The baseline performance of the binary classification is 50%. An average of 19.23% of improvement in the classification suggests that the voice quality decision is reflected to some degree in the PLPC features, which in turn suggests that we can conduct a PLPC-based speech recognition experiment utilizing the

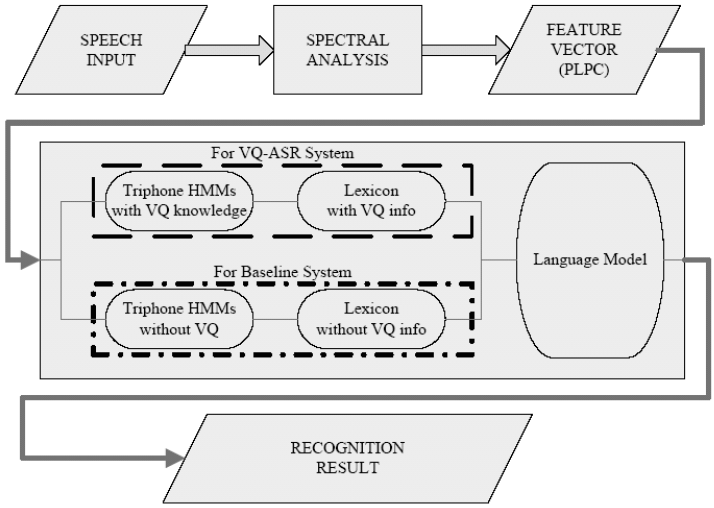


Figure 1: *The general automatic speech recognition architecture used in the baseline system and the voice quality dependent system*

voice quality information.

# 4 Voice quality dependent speech recognition

Given an acoustic signal $A$, the goal of a speech recognition system is to find the sequence of words $\hat{S} = \arg\max_S p(S|A)$. The general automatic speech recognition architecture is shown in Figure (1). Acoustic phonetic and phonological properties of speech sounds are represented by the HMM models, which usually have three emitting states, each having transitions either to itself or to the state on the right. The observation distribution of each state is multi-mixture Gaussian. Each HMM model corresponds to a triphone (e.g., "r-ih+k"), which describes allophonic variation by setting the acoustic properties of a given phone as a function of both the preceding and following phones.

## 4.1 Baseline system

We build a triphone-clustered HMM-based speech recognition system as the baseline system. In this system, we use HTK (Young et al., 2005) to cluster and tie the counterpart states, such as state 2 of "r-ih+k" and "s-ih+k", in some allophones

among the triphone HMM models according to a phonetic binary clustering tree. Every triphone unseen in the training data is synthesized by tying the states of the HMM to three states, chosen by the clustering tree, from seen allophones. Transition matrices of all allophones are tied together. Finally, Gaussian mixtures are incremented and parameters are further estimated to increase the number of mixture components in the output distribution of all states.

## 4.2 VQ-ASR 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.

**Inclusion of Voice Quality Information**: We use HTK to obtain the phone-aligned transcription from the Switchboard word transcription and wave files. This phone-aligned transcription is aligned against the voice quality label sequences given by the *frame voice quality decisions* described before. 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 it.

**Recognition Dictionary with Voice Quality Information**: 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. For example, for "bat b+ae b-ae+t ae-t" in the baseline system dictionary, the dictionary in VQ-ASR system should have two entries "bat b+ae b-ae+t ae-t" and "bat b+ae0 b-ae0+t ae0-t", where "0" represents the "creakiness label".

**Reduction of the Number of Parameters**: The number of triphones increases dramatically as the creakiness label "0" can be attached to the central phone and one or both of the neighboring phones, for each triphone. To reduce the number of parameters, we treat the triphones with different voice quality setting, e.g. "b-ae+t" and "b-ae0+t", as allophones of the same root monophone; both of these triphones are included in the same decision tree by the triphone clustering process. 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.

# 5 Experimental result

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. These systems are designed to identify the impact of voice quality dependency, and as such we do not compare our systems to full systems trained on much larger amounts of data (e.g., Luo and Jelinek (1999); Sundaram et al. (2000)). 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. 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 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. The number of mixtures in the HMM states are in the first column. %Correct is equal to the percentage of the reference labels that were correctly recognized. Accuracy is more comprehensive measure of recognizer quality that penalize for insertion errors.*

| Mixture | Baseline | | VQ-ASR | |
|---|---|---|---|---|
| | % Correct | Accuracy | % Correct | 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 |

# 6 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 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. 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


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.

Gerratt, B., Kreiman, J., 2001. Towards a taxonomy of nonmodal phonation. *Journal of Phonetics* 29, 365–381.

Godfrey, J., Holliman, E., McDaniel, J., 1992. Telephone speech corpus for research and development. In: *Proceedings of ICASSP*. San Francisco, CA.

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

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.

Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America* 87 (4), 1738–1752.

International Telecommunication Union (ITU) Standard G.711, 1993. Pulse code modulation (pcm) of voice frequencies.

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.

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*. pp. 353–356.

Maddieson, I., Hess, S., 1987. The effects of f0 of the linguistic use of phonation type. *UCLA Working Papers in Phonetics*.

Markel, J. D., 1972. The sift algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics* 20 (5), 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.

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*.

Taylor, P., 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107 (3), 1697–1714.

Yoon, T.-J., Cole, J., Hasegawa-Johnson, M., Shih, C., 2005. Acoustic correlates of non-modal phonation in telephone speech. *Journal of the Acoustical Society of America* 117 (4), 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* (for htk version 3.3). Tech. rep., Cambridge University Engineering Department, Cambridge, UK.