Tonality and its Application to Perceptual-Based Speech Enhancement

Jessica A. Rossi-Katz[1] *& Ajay Natarajan*[2]

Department of Speech, Language and Hearing Sciences[1], Department of Electrical Engineering[2]
University of Colorado, Boulder, USA
[rossija,nataraja]@colorado.edu

Abstract

In calculating the auditory masking threshold (AMT), it is necessary to adjust the excitation pattern of the clean signal to account for the asymmetry of masking between tone-like and noise-like signals. One approach that estimates the asymmetry of masking is the signal's tonality. The tonality assumptions were originally formulated for use in MPEG-4 audio coding. Consequently, they may not be appropriate for speech and hearing aid applications. There is some justification in using tonality-offset with voiced signals due to their formant structure. In this study, we differentially process the signal given a priori knowledge of its acoustic/phonetic characteristics (i.e. voiced, unvoiced and silence). Objective quality measures (e.g. segmental SNR and Itakura-Saito distortion measures) were used as yardsticks to compare schemes that include tonality-offset with those that do not. The segmental SNR shows an overall improvement of 1.1 dB for the tonality-offset scheme. Subjective quality measures based on an established rating scale were obtained from six listeners with normal hearing. Preliminary results indicate that schemes without tonality-offset are preferable to those with tonality-offset. One noted exception, though, is with the background noise attribute, where there is more than a 20% improvement in the tonality-offset scheme relative to other conditions. We present possible areas of improvement for implementing tonality-offset calculations in the future.

1. Introduction

A chief goal in designing hearing aids is to minimize the deleterious effects of competing noise on a desired signal. A majority of these noise suppression schemes use mathematically-based criteria (e.g. signal-to-noise estimates). However, their correlation with perceptual properties of the auditory system is limited. Recently, an enhancement scheme based on the masking properties of the auditory system was evaluated in listeners with normal hearing [1] and subsequently tested in listeners with hearing loss [2]. For both groups of listeners, auditory masked threshold-based noise suppression (AMT-NS) resulted in improvements in intelligibility and quality for some, but not all conditions. This paper serves as a follow-up to recent investigations in our laboratory and considers how components of the AMT-NS algorithm can be refined. Specifically, we consider the validity of adjusting the masked threshold based on the tonal characteristics of the speech signal.

It has been well-established that noise masks tones more effectively than tones mask noise [3]. Researchers have suggested that the bandwidth and temporal characteristics of the target and masker contribute to this asymmetry. Excitation patterns (EP), which represent the output of the auditory filters, are intrinsically associated with auditory masking [4]; if the excitation pattern (EP) of a target signal falls below that of a masker, the target stimuli is no longer audible. Coding applications have used these properties to compress audio and suppress noise [5]. Specifically, the EP is calculated by convolving the basilar membrane spreading function with the critical band densities. The EP is adjusted in accordance with the notion that tones and noise are asymmetrical maskers. This adjustment is the "offset" term. For a tone masking a noise, the EP is reduced by a factor of 14.5+i dB where i is the critical band number. For the converse condition (noise masking a tone), the EP is reduced by factor of 5.5 dB across critical bands. These values are based on results from [6] and [3] respectively. Both calculations are scaled, based on the degree to which a signal is noise-like versus tone-like (tonality), by computing a spectral flatness measure (SFM). The SFM is the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum. A SFM approaching 1 indicates that the signal is tone-like; a SFM approaching 0 indicates that the signal is noise-like.

Traditionally, asymmetry of masking has been investigated between pure tones and noise ([3], [7], [6]). Gockel et al. (2002) [8], parametrically investigated the asymmetry of masking between complex tones and noise, given that complex tones are more common in speech and music. Even with identical excitation patterns, there were large differences in the extent to which complex tones and noise were "mutual maskers". These differences varied with both overall level and phase of the constituent harmonics. Based on their results, Gockel and colleagues concluded that differences in masking efficiency between complex tones and noise varies as a function of level. Audio coders often do not incorporate level dependencies in their asymmetry of masking calculations and may underestimate the offset term, thereby overestimating the AMT. Certainly, Gockel et al.'s research, indicates that level dependency should be taken into consideration in future applications of the AMT.

Speech has both voiced and unvoiced segments. The formant structure of most voiced speech resembles the harmonic structure of a tonal signal. This is a key rationale for calculating tonality and offset terms for voiced segments and modifying these terms for unvoiced segments. In the current study, we evaluated whether AMT-NS schemes that employed tonality and offset calculations are more favorable than schemes which did not incorporate these calculations.

2. Algorithm Development

The flow chart of the algorithm is shown in Fig. 2. The noisy speech is broken down into frames and shaped using a Hamming window and a time-to-frequency transformation is applied using an FFT. An estimate of the clean spectrum is found using the GMMSE [9] algorithm. The Auditory Masking Threshold

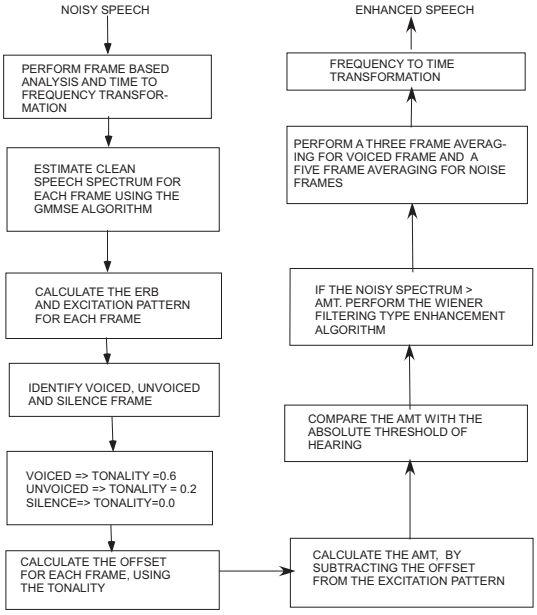

Figure 1: Flow chart of the enhancement algorithm

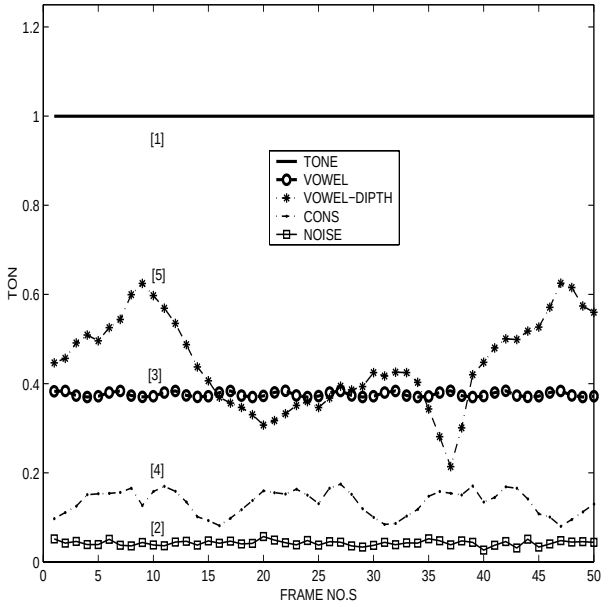

Figure 2: Tonality for different speech segments

(AMT) is found from the clean spectrum. Calculation of the AMT can be broken down into 4 steps

1. Do sub-band processing with the center of each band equal to the center frequency of each auditory filter. Auditory filters are represented using their equivalent rectangular bandwidth (ERB).
2. Calculation the excitation pattern for each sub-band using the Moore-Glasberg spreading function [10].
3. Calculate the offset term, based on the tonality of the speech waveform for each sub-band.
4. Subtract the offset from the excitation pattern and compare with the absolute threshold of hearing.

Next, the noisy power spectrum is compared with the AMT. If the noisy power spectrum is greater than the AMT, the signal is enhanced using an Wiener filtering operation [1].

2.1. Tonality Calculation

Fig. 2 illustrates the tonality for 1) a pure tone, 2) a noise, 3) the vowel /EY/, 4) the consonant /SH/ and 5) the vowel/diphthong pair /W-EY/. The tonality (ton) is measured using the ratio of the geometric mean (GM) of the signal and the arithmetic mean (AM) of signal, known as the spectral flatness measure (SFM). The equations for calculating tonality are shown below:

$$SFM(i)(dB) = 10\log_{10}(\frac{GM}{AM}) \quad (1)$$

$$ton(i) = \min(\frac{SFM}{-60}, 1) \quad (2)$$

where i is the sub-band number. From Fig. 2, it can be seen that the tonality of a pure tone is 1 and is close to 0 for noise-like signal. The tonality of the consonant is around 0.2 and around 0.4 for the vowel. We found interesting results when comparing the tonality of the vowel /EY/ with the that of the vowel-diphthong combination /W-EY/. It can be seen that the tonality of /W-EY/ varies between maximum of 0.65 and a minimum of 0.2. One would assume that the tonality measure for a vowel-diphthong combination would be less than that of a pure vowel. However, the SFM weighs the tonality of the present frame with the frame immediately preceding it and/or immediately succeeding it. If a vowel is preceded by a silence frame, the geometric mean of the signal in the transition frame decreases and the tonality of the transition frame increases. Another interesting finding, is that the tonality of vowels is rather low even though vowels have a formant structure that resembles the harmonic structure of a tonal signal. Due to these ambiguities in tonality measurements, we pre-set the tonality of the signal given a-priori knowledge of the speech file. Using the transcriptions of 10 TIMIT speech files, we set the tonality of voiced sections to 0.6, unvoiced/consonant sections to 0.2 and for silence frames we set the tonality to 0.

2.2. Offset Calculation

Johnston [5] used a linear equation for calculating the offset term in voiced frames. For this evaluation, we set the offset term to a flat value of 18 dB up to the first formant and linearly increased the term for frequencies above the first formant.

$$O_b(i) = ton(i)(18) + (1 - ton(i))6dB. \quad (3)$$

up-to the first formant and

$$O_b(i) = ton(i)(18 + i) + (1 - ton(i))6dB. \quad (4)$$

for all frequencies greater than the first formant, where ton(i) is equal 0.6

The constants in the equation are based on recent work by the Advanced Audio Coding group [11]. This sets the masking value of the vowel highest at the first formant location. This was not the case in either [6] or [5].

Consonants have low energy in the higher frequencies and are more noise-like at lower frequencies. To incorporate this idea, we set the offset term for consonants to a flat value of 6

dB for all frequencies up to 1.5 kHz and used a linear term like that in Eqn. 4, with ton(i) equal to 0.2, for all frequencies in the range of 1.5 kHz to 4 kHz.

We set the offset term for noisy frames to a flat value of 6 dB. This is same as that in Eqn. 4 with the ton(i) set to 0.

2.3. Frame Averaging

After calculating the AMT and performing noise suppression, we took a five-frame average of noisy frames and a three-frame average of voiced frames. This eliminates any irregularities and tonal structures in isolated frames. We performed the enhancement for 10 files given prior knowledge of the speech. In real-time applications, we can automate the voiced, unvoiced and silence frame detection by comparing the total energy in each frame with both low and high frequency energy.

3. Algorithm Evaluations

Ten single sentences from the TIMIT database were selected (5 male & 5 female). Each sentence was degraded with both communication (FLN) and large crowd room noise (LCR). The signal-to-noise ratio (SNR) was set at 5dB. The spectrograms of the noise types are shown in figure 3.

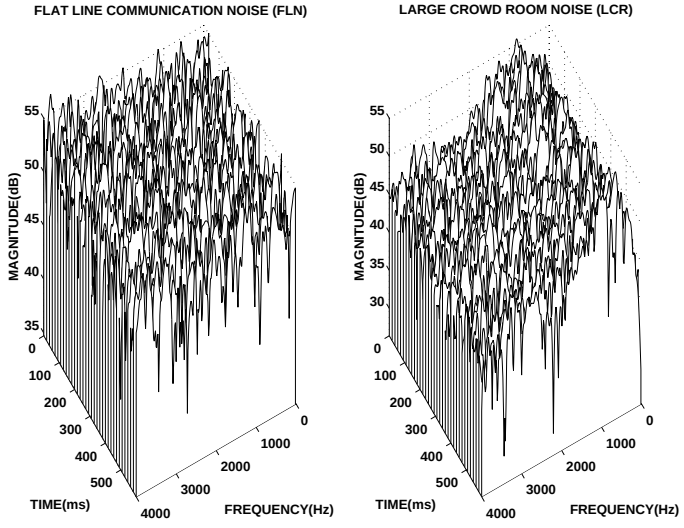

Figure 3: Time versus frequency spectrograms for different noise types: (a) Flat Channel Noise (FLN); (b) Large Crowd Noise (LCR)

Plots of clean speech, degraded speech and enhanced speech using both the no-offset (NOF) scheme and the offset scheme (OFF) are shown in Fig. 4. The plots are for the sentence, *"In wage negotiations the industry bargains as a unit with a single union"*. The shapes of the noise power spectrum, clean power spectrum as well as audible masked thresholds for NOF and OFF schemes for the unvoiced fricative /SH/ are shown in Fig. 5.

3.1. Objective Qualitative Measure

The results comparing the segmental SNR and Itakura-Saito (IS) distortion measure between 10 (about 4300 frames) degraded files (DEG), 10 enhanced files with no offset (NOF) and 10 enhanced files with offset (OFF) are shown in Table 1. Both

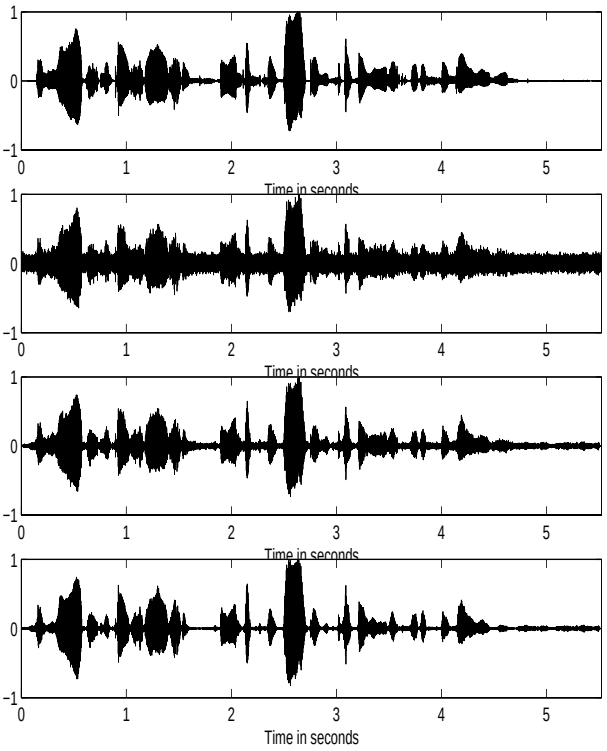

Figure 4: [Time waveforms for a single speech file] (a) Clean Spectrum (b) Degraded Communication Noise (FLN) at 5dB SNR (c) Enhanced speech waveform using no offset (NOF) (d) Enhanced speech waveform using an offset (OFF).

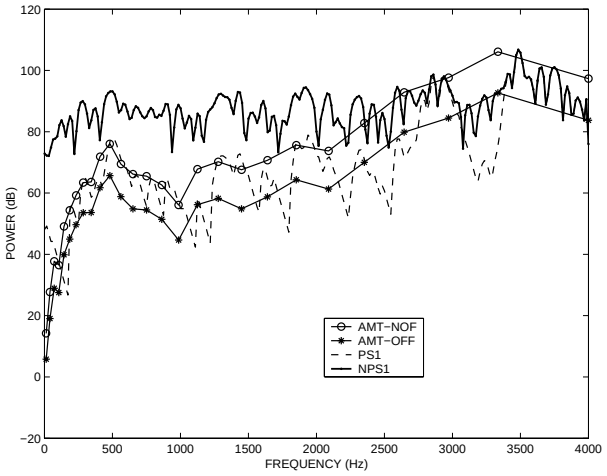

Figure 5: AMT NOF,AMT OFF, Noisy Spectrum (NPS) and Clean Spectrum (PS) of an unvoiced fricative /SH/.

enhancement schemes show significant improvements in SNR, and intuitively, the scheme with offset shows a greater improvement. However, the IS distortion measure in the OFF scheme is extremely high. By including an offset, the AMT is set lower and enhancement is higher. We also noted that enhancement is done about 58% of the time with the NOF scheme and about 90% of the time with the OFF scheme. This is the average for 10 files over two iterations.

Table 1: Objective quality comparison between the DEG, NOF and OFF schemes

Noise	DEG → NOF → OFF	DEG → NOF → OFF
	SNR	IS
FLN (5dB)	-1.807 → 2.227 → 3.379	3.206 →1.889 →10+
LCR (5dB)	-1.379 → 0.820 → 2.270	1.769 →1.445 →05+

3.2. Subjective Quality Ratings

This section introduces the testing protocol and compares the results of the subjective quality ratings between the DEG, NOF and OFF schemes. Six listeners with normal-hearing participated in this study. All listeners reported normal-hearing sensitivity and no significant history of ear disease. Listeners were tested individually in a double-walled sound booth. The test session typically lasted 30 minutes. For stimuli presentation, the digitally stored speech-in-noise stimuli went through a digital-to-analog converter, a 4000 Hz anti-aliasing filter, an attenuator and a headphone buffer. Finally, the stimuli were presented monaurally to the right ear of each listener through a TDH-49 earphone. All stimuli were presented at an RMS-equalized level of 65 dB SPL.

The categorical rating scales used for the quality ratings are the same as those used by Neuman et al., [12] and are similar to those developed by Gabrielsson et al. [13]. A 10-point scale was used to obtain ratings on five different stimulus attributes: clarity, pleasantness, background noise, loudness and overall impression. Listeners used a written response form to record their ratings. For each condition, participants listened to a block of 10 TIMIT sentences and rated each attribute using the 10-point scale. The order of the 10 sentences was randomized.

Table 2: Average quality ratings for DEG, NOF and OFF schemes for both noise types

	LCR			FLN		
	DEG	NOF	OFF	DEG	NOF	OFF
Clarity	5.5	5.78	4.2	3.94	4.33	4.0
Pleasantness	5	4.78	4.67	3.56	4.5	4.22
Background Noise	6.83	5.67	5.22	8.17	6.39	5.11
Loudness	4.89	5.33	4.67	4.67	4.94	3.83
Overall	5.0	5.33	4.22	3.67	4.06	3.97

The subjective quality rating are summarized in Table 2. For the background noise attribute, higher values are less favorable. For both noise types, listeners rated the NOF and OFF enhancement schemes significantly better than the degraded scheme; the OFF scheme was rated more favorably than the NOF scheme. On the other attributes, NOF was the highest ranked scheme, better than either OFF or DEG. The ranking of OFF and DEG varied between attributes and noise types.

4. Conclusions & Future Work

Preliminary results indicate that the NOF scheme was more favorable than both the OFF and DEG schemes. These results suggest that our initial offset measurements need fine-tuning. One reason that listeners may prefer the NOF or DEG schemes over the OFF scheme is due to the fact that inaccuracies in offset predictions affect later stages of processing; noise suppression algorithms also use AMT estimates when performing Wiener filtering. Due to increased noise suppression, speech signals are distorted in the OFF scheme. Consequently, listeners may be attuned to these distortions and the potential benefits of adjusting the masked threshold are negated.

A short-term follow-up study should evaluate intelligibility measures between the DEG, NOF and OFF schemes. These results may influence how the OFF scheme is modified in future studies. Detailed fine-tuning of the offset parameters must be done and take into consideration the level dependencies in masking as reported in [8]. Investigations into the theoretical underpinnings of asymmetry of masking will likely influence the estimations of the AMT. Certain schemes may be most beneficial when implemented with particular types of background noise. Future work should consider developing and evaluating a codebook that classifies the effects of processing schemes on different noise types.

5. References

[1] D. Tsoukalas, J. Mourjoupoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech & Audio Proccessing*, vol. 5(6), pp. 497–514, Nov 1997.

[2] K. Arehart, J. Hansen, S. Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," *Speech Communications*, 2003. In press.

[3] R. Hellman, "Asymmetry of masking between noise and tone," *Perception and Psychophysics*, vol. 11, pp. 241–246, 1972.

[4] B. Moore, *Perceptual Consequences of Cochlear Damage*. Oxford Psychology Series, 1995.

[5] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select Areas Commun*, vol. 6, pp. 314–323, 1988.

[6] M. Schroeder, J. Hall, and B. Atal, "Optimizing digital speech coders by exploiting the masking properties of the human ear.," *JASA*, vol. 6, no. 66, pp. 1647–1652, 1979.

[7] J. Hall, "Auditory psychophysics for coding applications," *CRC/IEEE DSP Handbook*, 1996.

[8] H. Gockel, B. Moore, and R. Patterson, "Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression," *Journal of the Acoustical Society of America*, vol. 111 (6), pp. 2759–2770, 2002.

[9] V. Radhakrishnan, "Speech enhancement based on generalized minimum mean square error estimation & masking property of the human auditory system," Master's thesis, University of Colorado, Boulder, August 2002.

[10] B. Moore and B. Glasberg, "Derivation of auditory filter shapes from notched-noise data," *Hear Research*, vol. 4, pp. 103–138, Aug 1990.

[11] I. JTC1/SC29/WG11, "Coding of moving pictures and audio-mpeg-2 advanced audio coding aac," *ISO/IEC 13818-7 International Standard*, 1997.

[12] A. Neuman, M. Bakke, C. Mackerise, S. Hellman, and H. Levitt, "The effect of compresion ratio and release time on categorical rating of sound quality," *Journal of the Acoustical Society of America*, vol. 103, pp. 2273–2281, 1998.

[13] A. Gabrielson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived sound quality of reproductions with different frequency and sound levels," *Journal of Acoustical Society of America*, vol. 88, pp. 1359–1366, 1990.