A New Acoustic Measure for Aspiration Noise Detection

Carlos Toshinori Ishi
Japan Science and Technology/CREST
ATR, Japan
carlos@atr.jp

Abstract
In this paper, we propose a new acoustic measure for detecting aspiration noise in vowels. The measure is an index of synchronization between frequency bands around the first and third formants. The measure is based on the principle that the vocal tract responses to the glottal excitation are synchronized between these frequency bands when aspiration noise is absent, and uncorrelated otherwise. Evaluation results show that the proposed measure can be used together with spectral slope measures for automatic detection of aspiration noise.

1. Introduction
Turbulent noise in speech can be produced by a sufficiently narrow constriction either in the vocal tract or at the glottis. When the constriction is in the vocal tract, fricative consonants are produced, while if the constriction is at the glottis, breathy voice or whispery voice is produced. According to Laver [1], in whispery phonation, the glottal constriction results from a triangular opening of the arytenoid cartilages of the glottis, while the vocal folds can vibrate independently. In breathy phonation, the constriction results from an incomplete closure of the vocal folds during its vibration.

Although breathy voice and whispery voice have distinct phonation settings, many researchers seem to use the term “breathy” for indicating both phenomena, since their auditory impression is of a turbulent noise due to an audible escape of air through the glottis. Breathiness is generally treated as a continuum that is difficult to separate into “breathy” and “modal”, whether sorting is based on perceived quality, acoustics, or the underlying glottal configuration [2]. The transition from “breathy” to “whispery” seems also to be part of an auditory continuum [1]. In the present paper we focus on the turbulent noise due to air escape only at the glottis (not in the vocal tract), and hereinafter, we will refer to it as “aspiration noise”. Fig. 1 shows the terminologies for the voice quality along a continuum and their relation with aspiration noise.

![Figure 1: Continuum showing relations between aspiration noise and voice quality.](attachment:figure1.jpg)

Aspiration noise in speech carries important linguistic and paralinguistic information, depending on the language. For example, a contrast between breathy and modal voicing among vowels is particularly common in many minor languages [3]. In [4,5], relationships between different phonation types and paralinguistic information like emotions and attitudes are reported. In [4], whispery voice was found in “fear” while breathy voice was found in “sad” voice. Correlations between synthesized breathy/whispery voices and perception of relaxed/stressed, sad/happy, and intimate/formal are reported in [5].

In the JST/CREST ESP Project, spontaneous speech is analyzed for Japanese, and possible use of breathiness for expressing manner or politeness in Japanese is reported [6,7]. Furthermore, aspiration noise is also frequent in the vowel portions of laughing speech. With the aim of doing automatic labelling of voice quality on a large speech database, in the present research, we focus on the automatic detection of aspiration noise.

2. Acoustic correlates of aspiration noise
For relatively small average glottal openings, as in modal phonation, the noise component is much lower in amplitude than the periodic component, while for relatively large glottal openings, voicing does not occur and only noise is generated (aspiration). For moderate glottal openings, as in breathy phonation, voicing may continue, but two changes occur: one is a high spectral damping in higher harmonics; the other is an increase in the amplitude of the turbulent noise because of the increased flow [8].

With respect to the spectral damping, a measure of spectral slope is proposed in [9] as an acoustic evidence of breathiness by taking the difference between the amplitude of the first harmonic ($H_1$) and the amplitude of the harmonic closest to the third formant ($A_3$). Another acoustic correlate of breathiness is reported in [10,11] for a parameter called $NAQ$, defined as the ratio of the peak-to-peak amplitude of the glottal waveform and the negative peak of the derivative of the glottal waveform. However, this measure requires inverse filtering of the speech signal, and seems to have problems to discriminate between breathy voice and nasalized vowels, since the amplitude of the fundamental in the inverse filtered signals is relatively strong in both cases. Both the measures ($H_1$-$A_3$ and $NAQ$) have drawbacks probably because they don’t take aspiration noise properties into account.

Regarding the aspiration noise factor, Klatt and Klatt [12] suggested a noise rating method employing the subjective ratings of visual observation of the irregularity of waveforms after band-pass filtering the signal around the third formant frequency ($F3$). In breathy phonation, the noise component in the $F3$ region becomes comparable to or stronger than the periodic component, and it is this significant noise component in the mid to high-frequency range that is believed to create the impression of breathy voice quality. Hanson [9] also applied this technique and both confirmed good correlation between raters.
2.1. Proposed method

Based on the subjective noise rating method proposed by Klatt and Klatt [12], we propose a method to quantify the degree of irregularities in the F3 region. When subjects evaluate the F3 band-passed signal looking at the waveforms, they seem to be comparing the shapes of the amplitude envelope of both signals to decide the irregularity rating. Based on this simple idea, we decided to estimate the amplitude envelope of the F3 band-passed signal, and quantify their similarity by taking their cross-correlation. However, instead of using the overall signal as the reference signal, we decided to use the signal filtered in a low frequency band, for two reasons: one is to avoid to take cross-correlation of two signals that always have a degree of correlation (since the overall signal contains the F3 band-pass signal); and the second reason is that the periodicity in low frequencies is less affected by aspiration noise in breathy voice [8].

Fig. 2 shows a simplified block diagram of the proposed method. The band-pass filtering is realized by taking the Fourier Transform of the input signal in a 32 ms frame length, and windowing (in the frequency domain) each frequency band. In order to avoid automatic formant extraction, we decided to set a fixed, broad range for each frequency band. For the F3 band, we set a range of 1800 ~ 4000 Hz, that is likely to contain F3 whether for male or female speakers. For the low frequency band, we set a range of 100 ~ 1500 Hz, since periodicity is less affected by aspiration noise in this region [8]. Although a band of 1000 Hz is suggested in [4], after preliminary experiments we decided to use a higher value (1500 Hz) in order to guarantee at least two harmonics even for very high fundamental frequencies (say, above 500 Hz), since the Hilbert envelope (applied in the subsequent block) of a pure sinusoid is a constant, and the periodicity would be missed. For simplicity, we will refer to the low frequency band as the “F1 band”, since this frequency range is likely to contain the first formant whether for male or female speakers. The band-pass filtered signals will be referred as $F1\text{wave}$ and $F3\text{wave}$.

![Figure 2: Simplified block diagram of the method proposed for aspiration noise rating.](http://www.limsi.fr/VOQUAL)

Next, the amplitude envelopes of the band-pass filtered signals are estimated for each frequency band. We first estimate the Hilbert envelope [13] of each frequency band, that provides an instantaneous amplitude of the signal. The amplitude envelopes ($F1\text{env}$ and $F3\text{env}$) are then obtained by smoothing the Hilbert envelopes of each band by using a Hanning window of 1 ms length. Such a smoothing process is necessary because the F1 and F3 bands have different bandwidths.

Now, the amplitude envelopes of F1 and F3 bands are cross-correlated to obtain an index of synchronization ($F1F3\text{syn}$) between these signals. The region for the cross-correlation calculation was set to the 25 ms center region of the frame, where the signals are expected to be stable. If the index of synchronization is low (the signals are uncorrelated), there is a high probability that the “noisy” components of the F3 band were produced independently of the glottal pulse. Therefore, the input signal is likely to contain aspiration noise.

Fig. 3 shows examples of the band-pass filtered signals and the respective amplitude envelope signals for several voice qualities of the vowel /a/, extracted from a female voice of our speech database. The energy envelope signals are normalized in the figure to allow a more suitable visual comparison. Fig. 3(a) shows the signals for modal voice. We can observe synchronized amplitude envelopes for $F1\text{wave}$ and $F3\text{wave}$, and also in the estimated envelopes $F1\text{env}$ and $F3\text{env}$. The index of synchronization $F1F3\text{syn}$ in this case is 0.75. Fig. 3(c) and (d) show examples of breathy voice and whispery voice. A clear periodicity can be observed in $F1\text{wave}$ for breathy voice, and a more unclear but still noticeable periodicity in $F1\text{wave}$ of whispery voice. However, no clear regularities can be observed in $F3\text{wave}$ for either breathy or whispery. Indeed, as expected, the amplitude envelopes $F1\text{env}$ and $F3\text{env}$ also reflect different shapes for both breathy and whispery, and the $F1F3\text{syn}$ values in both cases are low (0.05 and 0.03). Fig. 3(b) shows an interesting characteristic for creaky voice. As creaky phonation is often realized in very low fundamental frequencies, sometimes only one glottal pulse appears in the analysis frame, as shown in Fig. 3(b). In this case, even though there is no periodicity in the analysis frame, we can observe a similar behavior between the envelopes $F1\text{env}$ and $F3\text{env}$, and consequently a high $F1F3\text{syn}$ of 0.87.

For the breathy voice example of Fig. 3(c), we can observe a strong fundamental component relative to the third formant ($H1\text{-}A3$ is about 30 dB). However, in the whispery voice example, $H1\text{-}A3$ is smaller (about 10 dB). This indicates that although breathy voice has a steeper spectral slope for the periodic component [9], $H1\text{-}A3$ will not reflect this fact if the aspiration noise component is too strong relative to the periodic component. It is possible that such a measurement reflects the degree of aspiration noise components relative to the periodic component, when $F1F3\text{syn}$ is low. Here, we define a measure called $A1\text{-}A3$ as the difference (in dB) of the strongest component in the range of 100 ~ 1000 Hz and the one in the 1800 ~ 4000 Hz range. This measure also avoids automatic formant extraction, and reflects approximately the difference of the energy between the F1 and F3 bands.

3. Evaluation of the proposed method

As speech data for evaluation, we used a dataset of 50 utterances selected from our database of natural spontaneous conversational speech spoken by a female adult (in her 30’s). These 50 utterances were selected in a way to cover several speaking styles and voice qualities, including breathy voice, whispery voice, creaky voice, harsh voice, pressed voice, and laughing voice. (These utterances are part of the voice material prepared for the voice quality workshop VOQUAL 03, and can also be found from the following link: http://www.limsi.fr/VOQUAL). Preliminary observations of the data indicated that aspiration noise in vowels appeared in breathy, whispery, harsh and laughing voices.
Perceptual experiments were conducted where 8 subjects were asked to localize and annotate the syllables where aspiration noise is perceived within each utterance. Syllables annotated by 3 or more subjects were counted as containing aspiration noise. The distribution of syllable nuclei (total number of occurrences; number of syllables where aspiration noise was annotated) for each vowel type was: /a/ (168; 21), /e/ (73; 10), /i/ (34; 3), /o/ (79; 4) and /u/ (34; 4).

The 50 utterances were manually segmented at the phonetic level. The acoustic measures extracted for all the frames of the vowel portions were used for evaluation. 847 frames were obtained in the 42 vowels where aspiration noise was annotated, and 4532 frames in the 345 vowels where aspiration was not annotated. Fig. 4 shows scatter plots of the estimated measures ($F1F3syn$ and $A1-A3$) for the frames of vowels where aspiration was perceived (a) and not perceived (b). The frames around the phonetic boundaries and the modal portions of the vowels where aspiration was perceived were excluded from the scatter plot in Fig. 4(a), so that a total of 410 frames are shown in the figure.

It is remarkable that the vowels with aspiration noise consistently have lower $F1F3syn$ values (as expected) and $A1-A3$ limited around 30 dB. However, a large amount of samples whose aspiration noise was not annotated also yielded low $F1F3syn$ values. One possible reason for that could be the presence of aspiration portions in the vowels not
annotated as aspirated, but it certainly would not be the main reason. Another important trend observed in the scatter plots is that low \( F1F3syn \) values tend to concentrate on high \( A1-A3 \) values. If \( A3 \) is much lower than \( A1 \), i.e., big \( A1-A3 \) values (say larger than 30 or 40 dB), it is possible that the noise components in the F3 region are not clearly audible. Therefore, when \( A1-A3 \) has big values, the estimated value for \( F1F3syn \) should not be used for the aspiration decision.

Preliminary evaluation of aspiration noise detection was conducted using a threshold of 0.4 for \( F1F3syn \), and a threshold of 25 dB for the \( A1-A3 \) parameter. Results showed correct detection of aspiration noise in most frames of all vowels whose aspiration noise was annotated, but aspiration was also detected in some frames of vowels where aspiration noise was not annotated, as expected from the big overlap shown in Fig. 4. However, if a constraint is established where more than 3 consecutive frames of aspiration noise evidence is necessary within the vowel for being considered as aspirated, the matching with the subjectively identified syllables increases to more than 90 %. An explanation for this constraint could be that humans need a minimum duration (say 30 ~ 40 ms) of aspiration to perceive the syllable or phrase as aspirated.

Detailed analyses were also conducted at the frame level to verify the sources of low \( F1F3syn \) values in the vowels not annotated as aspirated. For most of the creaky frames, high \( F1F3syn \) values were observed. Low \( F1F3syn \) values were observed mainly in the phonetic transitions, especially from/to fricatives and plosives, since these phones also contain turbulent noise. This implies that \( F1F3syn \) values in the transition portions should be avoided in the decision of aspiration in vowels. Low \( F1F3syn \) values were also observed in some frames of modal phonation due to different shapes in the amplitude envelopes even though they were synchronized. Additional information of periodicity measurement in both bands could also be useful to avoid wrong decisions of aspiration in these segments.

With respect to the \( A1-A3 \) measure, in the present work, we used it as an approximation for measuring the difference of the energy between the two frequency bands. However, the strongest component in the F1 band (A1) could cause very different influences on the higher frequency bands; for example, a strong harmonic component around 100 Hz would have much less influence on the F3 band components, compared to a harmonic component around 800 Hz with the same strength. The use of auditory properties could be useful for deciding the thresholds for audibility of the F3 band components. This subject is left for future work.

4. Conclusion

A new acoustic measure (\( F1F3syn \)) was proposed for detecting the presence of aspiration noise in vowels. The measure reflected well the disconnection between the periodic and noisy components present in different frequency bands in vowels with aspiration noise. Evaluation of the proposed method showed its potential use for aspiration noise detection, but false detection occurred in some portions of vowels without aspiration noise. Investigations are being conducted to verify more deeply the source of these errors and to resolve these problems.

5. Acknowledgements

I thank Nick Campbell and JST/CREST for supporting this work. This work is a fruit of invaluable discussions with Parham Mokhtari (ATR), Hartmut Fittzinger (Univ. of Munich), and Mika Ito (Univ. of Edinburgh & Dublin). I also thank all the ESP group people who helped in the perceptual experiments, with special thanks to Minako Kimura.

6. References