Automatic Detection of Speech Truncation and Speech Rate

Chung Ting Justine Hui, Teh June Chin, Catherine Watson

Department of Electrical and Computer Engineering, University of Auckland

Abstract

Speech intelligibility can be affected when the speaker either speaks too fast or have portions of their speech omitted during the transmission process. This paper examines methods to measure the speech parameters that can be used to detect truncation in digital speech signal and to estimate the speech rate by utilising phonetic information in speech. Speech truncation was determined using techniques in the time and frequency domains, where the methods complement one another to estimate the likelihood of truncation. Speech rate was measured in terms of the number of syllables in an utterance and categorised into slow, normal and fast speeds.

Index Terms: speech truncation, speech rate, speech processing

1. Introduction

Successful communication requires speech to be conveyed in an intelligible manner. However, meanings can be easily misunderstood when part of the speech is omitted during the transmission process or when the speaker is speaking too quickly. These factors are especially important when communicating through devices like phones or hand-held radios used by emergency services where listeners can only rely on the transmitted signal to understand the messages being conveyed.

This paper proposes phonetically motivated techniques to provide feedback to users of hand-held devices the likelihood of truncated incomplete speech, where the utterance is cut off abruptly, and to categorise a given recording into slow, normal and fast speech. Obtaining a measure for these parameters will make both speaker and listener more aware of the possibility of speech degradation, especially in high stress emergency situations. The approach that was taken in this study heavily relies on the concept of grouping phonemes according to their different manners of articulation.

Figure 1: Sonority hierarchy

vowels → approximants → nasals → fricatives → affricates → stops

Within an utterance some sounds are more prominent or sonorous than others. The more prominent the sound, the more energy it has. Figure 1 shows how these sounds can be set up in a sonority scale from most sonorous to least [1]. These characteristics were used to develop the proposed methods for measuring speech truncation and speech rate.

Most automatic speech recognition (ASR) studies assume the utterance is complete with a definite start and end point in the speech [2] and therefore detecting truncation has not be considered very much. This process of locating the start and end of speech involves voice activity detection (VAD) [3] which is the process of separating the input signal into speech and non-speech signals and some of the common approaches include making use of fundamental frequency, root mean square energy (RMSE) and zero crossing rate [4]. While these methods do not solve our problem directly, the techniques used to obtain certain speech features, such as RMSE, can be applied to solve our problem.

On the other hand, measuring the speech rate has been studied thoroughly to improve ASR. Most literature defines speech rate as either the number of phones per second, or the number of syllables per second [5]. Since vowels are typically higher in energy than unvoiced consonants many of the studies concentrate on some sort of peak counting algorithms. This is the approach we have chosen to adopt and will be explained in detail in the following sections.

In the remainder of the paper, two separate experiments are presented. Section 2 describes the speech truncation experiment followed by its results, followed by speech rate in Section 3. We then conclude in Section 4.

2. Speech Truncation

2.1. Methodologies

Phonetically motivated techniques were chosen to obtain acoustic information from the speech signal to solve our problem. Our algorithms is then tested with the Australian National Database of Spoken Language (ANSDOSL) of 1000 English utterances that are 4-5 second duration, spoken by five male speakers [6]. The speech signals in this database are fully labelled at the phonetic, phoneme, syllable and word levels. This database allows us to solve our problem using a phonetics approach, making use of the specific characteristics of individual phonetic groups.

2.1.1. *Change of Rate of Energy Over Time*

If we look at the energy trace of a speech signal with a clear start and end, we can observe the signal increasing and decreasing gradually over time. From this observation, we can also expect a truncated speech signal to display a steeper gradient in its energy trace than that of a non truncated signal. This method, which is referred to as the RMS method, makes use of this difference in gradient by taking the RMS energy trace of the speech signal to determine whether or not the signal has been truncated. We can observe a clear difference in the change of rate in Figure 2a and Figure 2b to determine the likelihood of truncation.

However, the RMS method is unable to accurately detect truncation when it occurs at a sound with low intensity and hence another method had to be implemented to account for this.

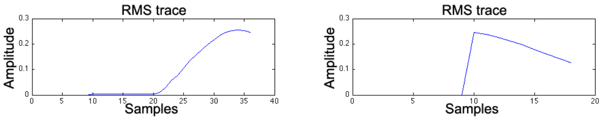

Figure 2: *RMS trace of recording to first onset of speech for a non-truncated signal(a) and truncated signal at 400ms into /a/ of the word 'thank'(b)*

2.1.2. *Overall Energy*

While we make use of the dynamic change of the energy in the signal , the amount of energy in the speech signal can also gives us an indication of truncation in the signal. This method, referred to as the SS method, looks at the spectral section by taking the Power Spectral Density (PSD) on the signal. The size of the area underneath the PSD determines whether or not the utterance is truncated.

As we can see in Figure 3a and Figure 3b, the energy measured for a non truncated piece of speech is considerably lower than that of a truncated piece of speech. The SS method is best used to determine whether a supposedly low energy segment in the time domain is actually still in the midst of an utterance. We can see a comparison of the RMS method and the SS method in Figure 4. While it would have been difficult for the system to determine a truncation has occurred by its change of rate, the amount of energy in the signal is considerably higher than that of silence, and thus we can predict some sort of presence of speech.

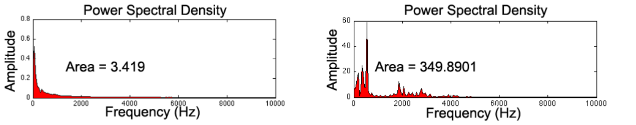

Figure 3: *Area under PSD at start of recording for an non-truncated signal(a) and truncated signal at 400ms into /a/ of the word 'thank' (b)*

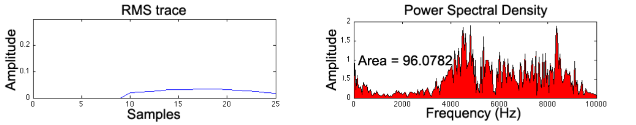

Figure 4: *RMS trace (a) and PSD (b) of the recording starting in the midst of an utterance (truncated 900ms into recording at the stop 't of the word 'it')*

2.1.3. Periodicity of Source Estimation

Linear Predictive Coding (LPC) models the production of speech as a result of the glottis, the vocal tract and the lips [7]. The excitation of the glottis characterises the intensity and fundamental frequency of the signal, while the vocal tract defines the formants of the signal. Removing the formants and attenuation by lip radiation from the original signal results in a residual signal for our analysis.

We assume that the signal outside the sample analysed is stationary, and therefore the residual signal of non truncated speech will exhibit an impulse train, where the peaks are distributed regularly with a gradual increase in magnitude [8] as marked by the asterisks in Figure 5a. This method makes use of this periodicity of the peaks in the residual signal to determine whether or not the signal has been truncated. For a voiced sound that is truncated, the signal will most likely start at a point offset from zero, thus creating a large difference between the original and estimated signals in the first few samples of the recording. This introduces a large peak at the front of the residual signal that disrupts the distribution of the impulses for the whole signal as displayed in Figure 5b.

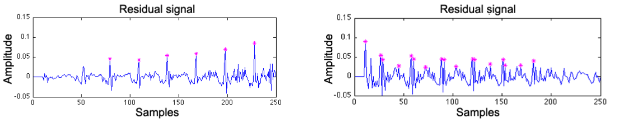

Figure 5: *Residual Signal for an non-truncated signal where peaks are distributed periodically (a) and truncated signal where peaks are distributed aperiodically (b)*

2.2. Estimating truncation

Each of the 1000 recordings in the database were analysed by incrementally truncating them at intervals of 50ms, creating a total of 79,938 instances of speech data,which includes truncation that occurred at silences before or after an utterance, or truncation that occurred at a particular phonemic sound. Each of these instances of truncation are represented by the vertical lines in Figure 6 where they were analysed with the RMS, SS and LPC methods.

2.3. Results

Figure 6 displays the methods that have detected a possibility of truncation at each 50ms instance. The green dots represent none of the methods have detected truncation, and each of the other colours, blue, red and yellow represent each of the methods, LPC, SS and RMS, having detected truncation accordingly.

As shown in Table 1, RMS performs least accurately in detecting truncation during an utterance, but performs most accurately in determining non-truncated speech, which acts as a control factor. This means that when we apply the method at a segment of silence, the method does not regard the instance as being truncated, which is the expected response. On the other hand, SS gives the best results for detecting truncation during speech, but it is the worst in accounting for silences. This means that the SS method would regard a truncation being present even during silences before or after speech in the signal. LPC stands in the middle between the RMS and SS method, where its utterance detection accuracy scores lower than SS but higher than RMS, and vice versa for the silences. Figure 7 divides the utterances into their phonemic groups to show how well the methods perform for different speech sounds. For high prominent sounds such as vowels, approximants and nasals, all three methods tend to perform fairly accurately. However while the SS method can detect truncation with a high probability during stops, fricatives and affricates, the RMS method performs quite poorly in detecting truncation during these sounds. By combining the three methods, we are able to evaluate the likelihood of truncation during an utterance, depending on how many methods have identified a possibility of truncation.

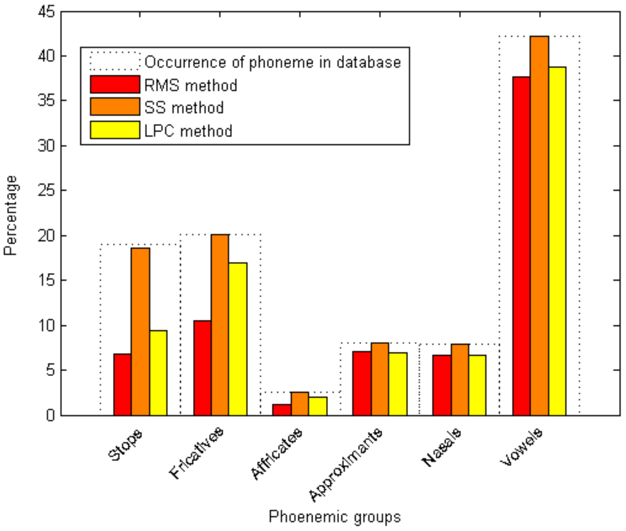

Figure 7: *How each method perform within the phonemic groups and their occurrences in the database*

2.4. Discussion

The SS method is the most sensitive where it has the highest error rate in detecting actual silences as truncated speech. This is because the SS method at times incorrectly regards an intake of breath or any insignificant background noise present as speech. On the other hand, the RMS method is the most conservative of the methods, where it can account for silences unlike the SS method. Therefore, when the RMS method deems a signal as truncated, we can be almost certain that the speech has been cut off. LPC is the most technical method in determining truncation because of the number of scenarios the method can produce. The LPC method also tends to be more unpredictable as to where it would fail. For example, the LPC method would fail when a truncation were to occur in such a way that reduces the error between the predicted and original signal, creating no peaks in the residual signal, regardless of the type of speech sounds.

Since each of the methods explained above has their own advantages and disadvantages, we propose to use them in parallel to complement one another in order to give an indication to the users whether or not truncation has occurred.

3. Speech Rate

For the purpose of this paper, we have defined speech rate as the number of syllables per second. Syllables are distinguished by the presence of vowels and therefore by identifying vowel occurrence

Table 1: *Accuracy of each method detecting truncation and silence*

Methods	Utterance	Silences
RMS	69.99%	98.17%
SS	99.51%	67.71%
LPC	80.71%	97.20%

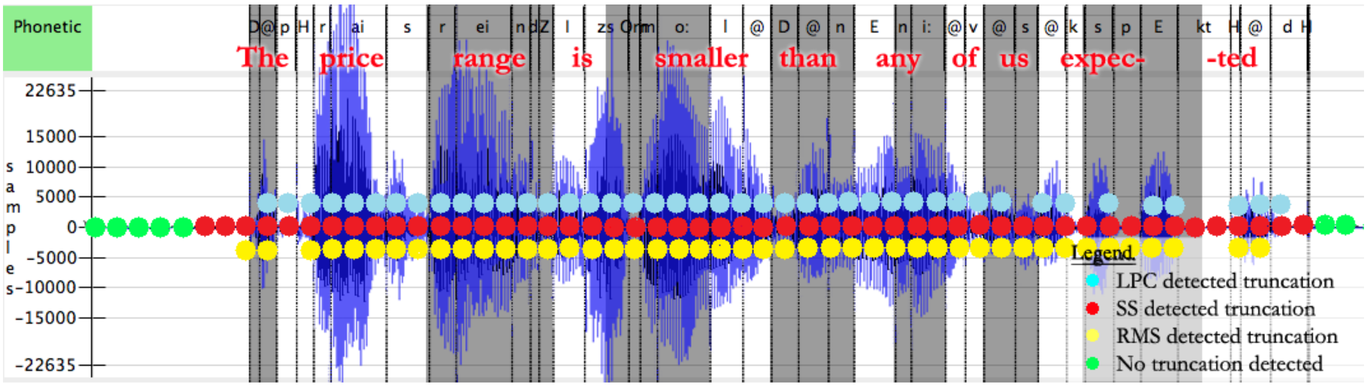

Figure 6: Example sentence with the methods performed at 50ms segments

we are able to count the number of syllables. Because vowels are always voiced, they tend to be higher in energy than other speech sounds in the speech stream [1]. By using an energy contour, we can identify the peaks as the location of the vowels, and as a result we can obtain the number of syllables in the utterance. Number of syllables per minute in the utterance is then categorised accordingly into the following speeds: slow, normal and fast.

3.1. Methodologies

3.1.1. Root-mean-square

Using an RMS trace of the speech signal, the vowels are identified from the peaks in the energy profile. The number of vowels are counted using a peak detection algorithm in MATLAB. Figure 8 shows the vowels in the utterance being highlighted underneath the RMS contour of the signal represented by the dots.

However, vowels are not the only sounds that have a high prominence in speech. Speech sounds such as nasals, approximants, and some higher energy fricatives such as /s/, also form a peak in the RMS trace.

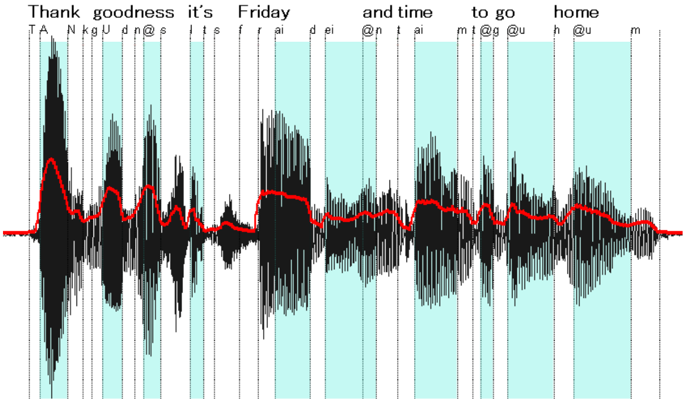

Figure 8: *RMS contour for speech rate estimation*

3.1.2. Zero-Crossing

Zero-crossing rate is the frequency of how much the signal crosses the zero line. This method is used to address the peaks contributed by the high energy fricatives and using this information, we can subtract these non-vowel peaks from the nuclei contour. Figure 9 shows the zero-crossing rate plot of the same speech signal as Figure 8. As we can see in Figure 8, the /s/ in the words “goodness” and “it’s” also contribute to the peaks in the energy contour, but at the same time, /s/ also has a relatively higher zero-crossing rate as shown in Figure 9. By using the zero-crossing method, we can eliminate the higher energy fricatives in the RMS trace.

3.1.3. Teager Energy Operator (TEO)

Other problematic sounds such as nasals and approximants exhibit vowel-like qualities, meaning that they are usually voiced and higher in terms of energy, and therefore creating undesirable peaks in the energy contour. To account for that, we have applied the TEO defined by Equation 1 [9].

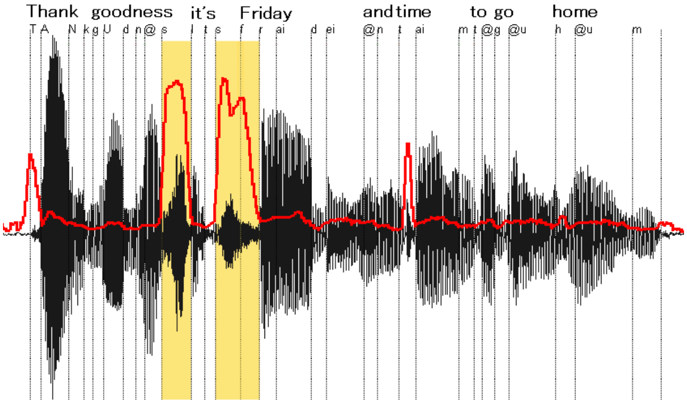

Figure 9: *Zero-crossing rate for speech rate estimation*

$$E_i = x^2[n] - x[n-1]\,x[n+1] \tag{1}$$

where E_i gives the running estimate of energy in the signal. After applying the TEO to the signal, an RMS trace of the TEO is calculated to compare with the RMS trace of the original signal. As we can see with the example on “annul” in Figure 10, while the RMS trace of the original signal has no distinct peaks to show where the vowels are located, the vowels can be clearly identified in the RMS trace of the TEO applied signal.

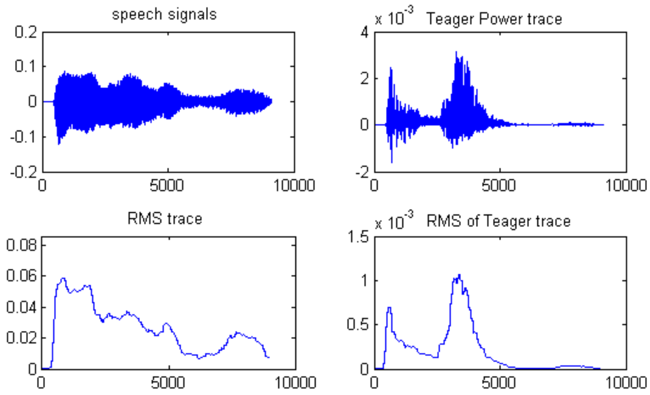

Figure 10: *TEO example on the word “annul”*

The signal in Figure 11 shows how the TEO can help in diminishing the extra peaks from the nasal nuclei. The sentence itself has 11 syllables, but the RMS trace has detected 13 syllables, with the two extra syllables contributed from the nasals in ’and’(/ənd/) and ’home’ (/həum/). On the other hand, in the RMS trace where the TEO has been applied, the peaks formed from the nasals are considerably diminished. However, we have found that there are instances when the TEO smooths out the signal too much and therefore the operator is used in conjunction with the unmodified RMS trace to estimate the number of syllables.

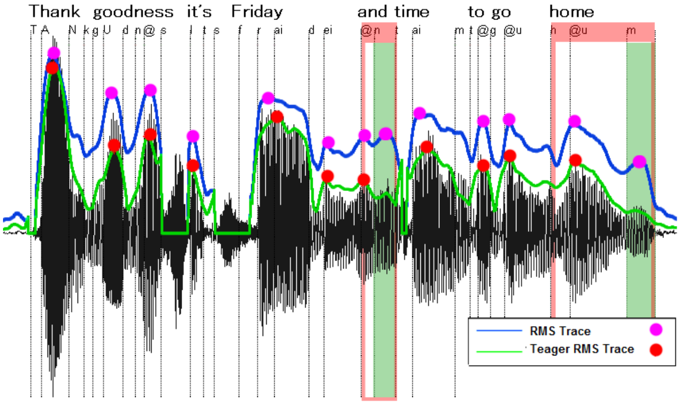

Figure 11: *Using a combination of both methods*

3.1.4. Categorisation of Speech Rate

The number of syllables per second is categorised into slow, normal and fast speeds to give a more meaningful feedback to the speaker, where 280 syllables per minute (spm) is regarded as the average speed for New Zealand English [10]. With our data taken from Australian English speakers, we have categorised the spm we have recorded into: slow speech being 250 spm or below, normal speech between 250 to 320 spm, and fast speech 320 spm or above, influenced by Tauroza's work on British English [11]. We have assumed that the number of syllables in a speech stream is proportional with time. Therefore we have simply multiply our results, which are in the unit of number of syllables per second, by 60 to calculate the speech rate as syllables per minute.

3.2. Results

The proposed methods were tested on the 1000 recordings in the database to estimate the number of syllables in the utterance. This estimated count was compared to the theoretical number of syllables supplied by the database. By looking at the differences between the detected number of syllables and the theoretical number of syllables, we are able to test how accurate our system performs in estimating the speech rate. Table 2 shows the number of utterances that has been miscounted with the number of syllables in absolute form, where the difference can either be an underestimation or overestimation. We then looked at how accurately our system can categorise the speech rate correctly. Table 3 shows the percentage of the prototype classifying the utterances into the speed categories comparing to the theoretical number. 8.8% of the time our system mistakenly classify fast or normal speech as slow, 11.6% of the time normal speech is mistakenly classified as slow or fast speech and 2.8% of the time normal or slow speech is mistakenly classified as fast.

Table 2: *Percentage of miscounting the number of syllables*

Syllable Count	0	1	2	3	4	5	6	7	8
Percentage	19.9	38	21.6	12.4	5.2	1.6	4	0.8	0.1

Table 3: *Percentage of speech categories from the database*

Speed	Theoretical %	Detected %	Difference %
Slow	6.6%	15.4%	8.8%
Normal	48.2%	36.6%	11.6%
Fast	25.2%	28%	2.8%

3.3. Discussion

For speech rate, because the TEO can respond quickly to the changes in both the amplitude and the frequency, it diminishes low frequency sounds [12]. According to Gimson, nasals and some approximants have a low frequency murmur [1]. We speculate that it is this characteristic of the sound that allows the TEO to successfully suppress these speech sounds, as shown in Figure 10, the /n/ and /l/ sounds.

While the TEO seems like a good solution, in some cases it underestimates the number of syllables from flattening out the waveform excessively. Using only the original RMS trace modified by the zero-crossing rate function, we can estimate speech rate within a margin of 4 syllable difference 96% of the time, comparing to 97.1% of the time when the TEO is applied. This means that the TEO only improves our algorithm by approximately 1%. However, we believe that there is more potential in what the TEO can do. In other studies, the TEO has been used to detect hypernasality in speech by using a difference between two Teager Energy profiles [13]. Therefore, more research can be done in applying the TEO as a speech signal processing tool to improve the accuracy of calculating the speech rate.

4. Conclusions

We have based our work around clean signals, which is impractical for real life situations. To address this issue, noise filters can first be applied to the signal, as well as utilising a more sophisticated VAD that takes noisy signals into account. In this paper, we showed that by addressing different speech sounds using their specific phonetic characteristics, we were able to estimate whether or not a speech signal has been truncated and categorise the speed at which it was spoken at. As a result, we have developed a platform that makes use of the linguistics information of a speech signal, offering a different way of addressing problems in this area of research. A trial version based on this study is currently being implemented to be used for emergency services.

5. Acknowledgement

We would like to thank Alan Murray from Tait Electronics for his guidance and support.

6. References

[1] *Gimson's Pronunciation of English*, 5th ed. E. Arnold, 1994.

[2] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. 3808–3811.

[3] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 965–974, 2005.

[4] K. Y. Luu, "Real-time noise-robust speech detection," Master's thesis, Massachusetts Institute of Technology, 2009.

[5] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters," in *12thAnnual Conference of the International Speech Communication Association*, Florence, 2011.

[6] *Australian national database of spoken language (andosl)*, 1999. [Online]. Available: `http://andosl.anu.edu.au`.

[7] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[8] C. D'Alessandro, "Spoken language processing," in, J. Mariani, Ed. Wiley, 2009, ch. Speech Analysis.

[9] H. A. Patil1 and T.K.Basu, "Identifying perceptually similar languages using teager energy based cepstrum," *Engineering Letters*, vol. 16, no. 1, 2008.

[10] M. P. Robb, M. A. MaClagan, and Y. Chen, "Speaking rates of american and new zealand varieties of english," *Clinical Linguistics and Phonetics*, vol. 18, no. 1, pp. 1–15, 2004.

[11] S. Tauroza and D. Allison, "Speech rates in british english," *Applied Linguistics*, vol. 11, no. 1, pp. 90–105, 1990.

[12] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified teager energy measurement," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1993, pp. 732–735.

[13] D. A. Cairns, J. H. Hansen, and J. F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear teager energy operator," 1996.