Modulation Spectrum for Pitch and Speech Pause Detection

Olaf Schreiner

DaimlerChrysler Research and Technology
and University of Göttingen, Germany
olaf.schreiner@daimlerchrysler.com

Abstract

This paper describes a new approach to the speech pause detection problem. The goal is to safely decide for a given signal frame whether speech is present or not in order to switch an automatic speech recognizer on or off. The modulation spectrum is introduced as a method to determine the amount of voicing in a signal frame. This method is tested against two standard methods in pitch detection.

1. Introduction

Automatic speech recognizers usually have the capability to classify frames as pause. So, why do we need to know in advance whether there is speech present or not?

Most of the current speech recognizers use Hidden Markov Models (HMM) to classify a portion of the speech signal. For every speech fraction entity (phonemes, syllables, words) a separate model is trained. Pauses are not treated different from any other speech sounds. Thus, the pause model is trained to whatever the typical pause in the training material sounds like. In the recognition case, the background noise may be different and the pause model may not match the background noise with the desired precision. In addition, the pause model is trained to the typical length of a pause in the training material. That may causes a pause of very different length in the recognition case to be less likely classified as pause, and, in effect, causing the recognizer to put out words where none have been spoken.

In command and control dialogs, one or more words spoken together make up a command utterance that is passed on to the dialog system. The dialog system interprets the utterance and initiates actions to be taken. Usually, between utterances there are significant pauses that can be used to separate the utterances from each other. Hence, if the recognizer is turned on upon the first word in the utterance and turned off at the first significant pause, this “take” very likely contains one utterance.

To accomplish this speech detection, one can make use of the fact that every syllable of a word contains vowels, i.e. voiced sounds. One way to detect vowels is to look for strong periodicities in the speech signal, as do the autocorrelation method or the cepstrum method. The current approach is even more restrictive and demands that a voiced sound is amplitude modulated.

2. Modulation Spectrum

Detection of voicing is strongly correlated with the detection of the pitch frequency. Langner et. al. [4] showed that in the human auditory system the perception of pitch is not mainly done by the frequency decomposition on the basilar membrane but by a further analysis in the auditory cortex. This analysis detects amplitude modulations in the given frequency channels.

Prior work has been done to model this analysis of amplitude modulations [7]. The easiest way to model the frequency decomposition would be a power spectrum accomplished by Fourier Transform. A time series of power spectra is also known as spectrogram. If the time window for the Fourier Transform is chosen short enough (< half pitch wave length), a time structure can be seen in the spectrogram. For voiced sounds this time structure is consistent through all frequency channels (see fig. 1).

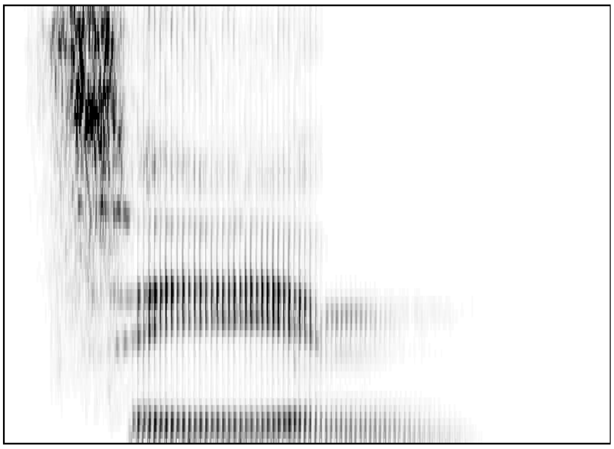

Fig. 1: Spectrogram of the German word “10” /tse:n/

This time structure (the vertical bars) repeats with the pitch frequency. Thus taking a power spectrum, again, of a frequency channel of the spectrogram yields a modulation spectrum for the frequency channel. The peak in this modulation spectrum will most likely be at the pitch frequency. Taking the modulation spectrum for each channel provides the Amplitude Modulation

Spectrogram (AMS, see fig 2b). The AMS can further be adapted to the human auditory system by using a logarithmically spaced frequency decomposition (e.g. by wavelet transform). But the latter is computationally very costly and shall thus not be further discussed.

According to the formant structure of a given speech sound the frequency channel containing the strongest modulation will vary. On the other hand, computing all frequency channels of the AMS can be critical in real time applications. Thus, it would be desirable to have a simpler spectrum of the overall amplitude modulation of the signal frame. Therefore, we first need a "power signal". An approximation of a signal that behaves like a frequency channel of the spectrogram is the Hilbert Envelope

$$h(t) = \left| s(t) + \frac{i}{\pi} \int_{\tau \neq 0} s(\tau)/(t-\tau)d\tau \right|$$

with s(t) being the time signal. Finally, the power spectrum of the Hilbert Envelope yields the overall "modulation spectrum" (fig 2c).

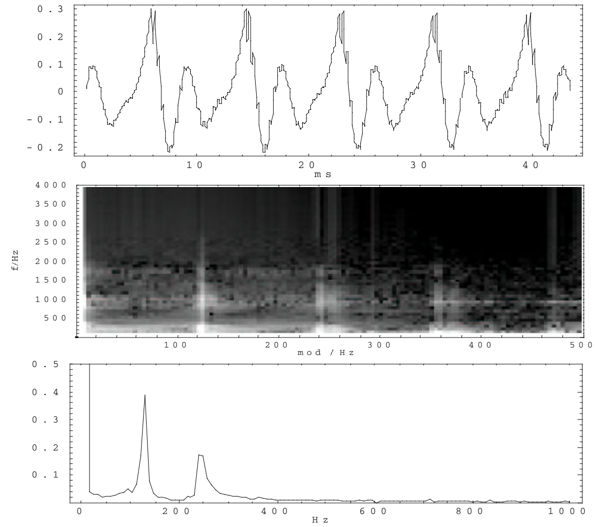

Fig 2: Male /e:/ – (a) waveform, (b) modulation spectrogramm, (c) modulation spectrum

3. System Overview

Computing the modulation spectrum by the time domain integral would still be too expensive. Therefore, we note that the Hilbert Envelope is the absolute of the analytic signal σ.

$$h(t) = |\sigma(t)|$$

The analytic signal itself can be obtained by Fast Fourier Transform (FFT) of the speech signal, setting the negative frequencies to zero and reverse FFT. Taking another FFT of the absolute of the analytic signal and finally the absolute of that yields the modulation spectrum.

To achieve a signal independent criterion for the strength of the laryngalization ("voicedness"), the modulation spectrum is normalized to the overall signal energy. The maximum of the normalized modulation spectrum finally yields the extent of voicing, which roughly lies between zero and one. A proper threshold for tagging a frame "voiced" is about 0.3. Dynamic Programming [6] is used to find the most likely chain of maxima. The search is restricted to valid speech frequencies (roughly from 50 to 500Hz). The modulation spectrum values are used as state probabilities. Since for physiological reasons the fundamental frequency can only change so much from a given time frame to the next [3], big jumps in pitch frequency are punished with low transition probabilities. A viterbi path length of 10 frames considerably improves detection precision [5].

The speech signal is sampled at 16 kHz. Window length for computing the analytic signal is 64ms (1024 samples).

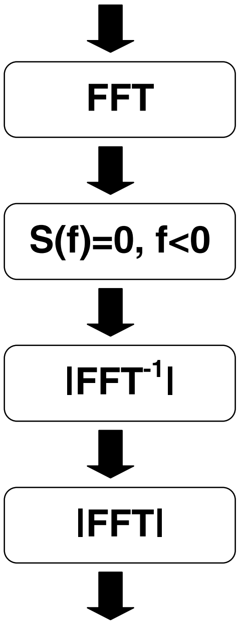

Fig. 3: Computation of the Modulation Spectrum

4. Extensions

The modulation spectrum does not only provide a peak at the fundamental frequency, but the whole modulation spectrum. Therefore, the modulation power may be distributed over a set of frequencies, namely the harmonics. Low male voices especially have a tendency to contain strong peaks at the double of the fundamental frequency (fig. 2c). Soft female voices tend to have a great number of small valued harmonics, stealing a considerable amount of power from the fundamental frequency peak. Noise, on the other hand, has a very flat Modulation Spectrum. Thus, the variance of the Modulation Spectrum contains some extra information about how 'spikey' the spectrum is. The fundamental frequency peak multiplied with the variance is then a safer criterion for a frame being voiced.

Vowels are usually longer than one 10ms time frame. Therefore, instationary noise can be reduced by taking the spectral minimum of three consecutive time frames in each frequency channel. To further reduce distortions each frequency channel is smoothed with an exponential time window.

5. Experiments

In the experiments, the detection of speech in a signal has been measured directly. The test involved the SPONTI database [1,2] of the University of Erlangen, Germany. The database contains about 1300 short sentences of about three to seven words, uttered in a spontaneous manner. Every sentence is hand labeled for pitch frequency and voiced/unvoiced.

Tests were performed for detection of syllables. Syllables are defined here in a very sloppy manner, meaning any block of voiced frames that a voiced / unvoiced detector delivers or that have been labeled so. A syllable is considered found, if there is any overlap between a detected syllable and the labeled syllable, this, again, being a very weak condition. The measurements dealt with labeled syllables that are not found by the detectors, the so called deletions. Also, bogus syllables, meaning detected syllables where none are labeled, were counted. A third test measured the relative syllable length to find out how much of the syllable onsets and offsets have been cut off by the detector. Additionally, as we deal with pitch detection algorithms, the precision of the fundamental frequency that comes as a byproduct has been measured.

All measurements were made at different noise levels, including 0dB, 3dB and no noise (>>20dB). Since the SPONTI database was recorded in a quiet office environment it does not contain any noise. Thus, noise at different levels has been added. The used noise was recorded in a Mercedes Benz sedan with no special noise sources (like open windows or rain) at an average speed of 120 km/h. Only voiced parts, i.e. parts of an utterance with at least half the maximum energy, were regarded for SNR determination. Noise was then added with constant energy throughout the whole utterance. For better discrimination, pre and post gaps have been disengaged in the experiments.

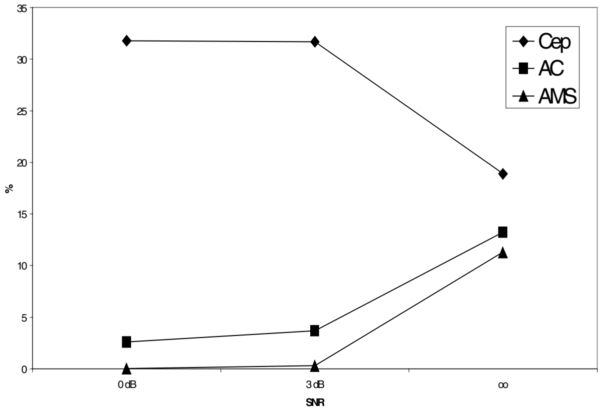

Fig. 4: Insertions of syllables

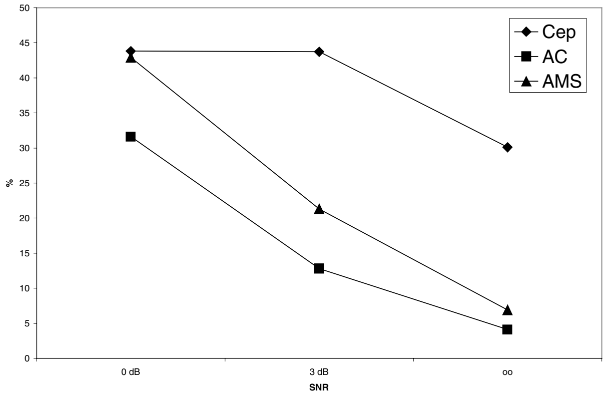

Fig. 5: Deletions of Syllables

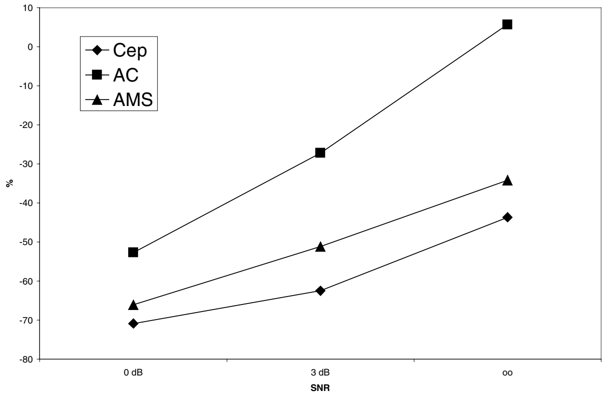

Fig 6: Relative length of syllables

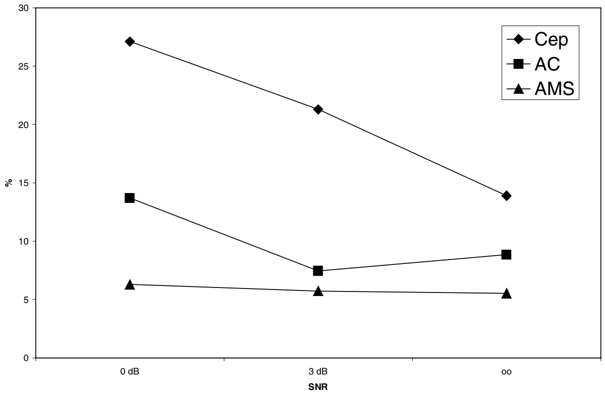

Fig. 7: Pitch Precision

6. Discussion

For Insertions of syllables (fig. 4), the modulation spectrum clearly beats the other algorithms at all noise levels. At a signal to noise ratio of 3dB, which is a realistic scenario for in-car applications, there are virtually no insertions (0.3%).

In terms of deletions (fig. 5), the autocorrelation turns out best, followed by modulation spectrum. This is mostly because unemphasized first or last syllables in a word may not have full laryngalization and are thus better detected by autocorrelation. If pre and post gaps are introduced, most of those affix syllables, missed by the modulation spectrum, can be rescued.

The syllable length (fig. 6) is a measure of how much is cut of at the beginning and the end of each syllable. Here again, the autocorrelation shows the best results. Again, weaknesses in this discipline can be fixed by pre and post gaps.

In pitch precision (fig. 7), the modulation spectrum shows best results, followed by autocorrelation. The cepstrum proves to be the weakest in all disciplines.

Modulation spectrum and autocorrelation each are best in two disciplines. For speech recognition purposes the winning disciplines of the modulation spectrogram seem more critical and the weaknesses in the other two can easily be overcome as described above.

7. Conclusions

The modulation spectrum has been introduced as a pitch detection algorithm. It has been tested for fitness in speech pause detection and compared to two standard algorithms in pitch detection, autocorrelation function and cepstrum. The modulation spectrum performs similar to the autocorrelation function, but slightly better in two critical categories.

8. Acknowledgements

We would like to thank the pattern recognition group at the University of Erlangen-Nuremburg, Germany, for kindly providing the test material.

9. References

[1] Batliner, A., R. Kompe, A. Kießling, E. Nöth, H. Niemann: *Can You Tell Apart Spontaneous and Read Speech if You Just Look at Prosody?*, in Rubio Ayuso, A.J. and López Soler, J.M. (eds.): *Speech Recognition and Coding. New Advances and Trends*, Springer, Berlin, NATO ASI Series F, 147, 1995, pp. 321-324.

[2] Batliner, A., W. Oppenrieder, H. Altmann: *Grammatisch-intonatorische Modus- und Fokusmarkierung in unterschiedlichen Registern gesprochener Sprache*, DFG-Abschlußbericht, Manuskript, München, 1992

[3] Hess, W.: *Pitch Determination of Speech Signals*, Springer, Berlin, 1983.

[4] Langner, G., H. Schulze, M. Sams, and P. Heil: *The topographic representation of periodicity pitch in the auditory cortex*, Proc.of the NATO Adv. Study Inst. on Comp.Hearing 1, pp. 91-97, 1998.

[5] Quast, H., O. Schreiner and M.R. Schroeder: *Robust Pitch Tracking in the Car Environment*. Proc. of Int. Conf. on Acoustics, Speech and Signal Processing, 2002.

[6] Rabiner, L.R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, in A. Waibel and K.-F. Lee (eds.), *Readings in Speech Recognition*, Morgan Kaufmann, San Mateo, 1990.

[7] Schreiner, O. and H.W. Strube: *Modulationsfilterung mit Fourierspektrogramm und Wavelettransformation*. Fortschritte der Akustik – DAGA, 2001