An auditory filter representation of phoneme boundaries

Bram G. Alefs

Institut für Schallforschung, Österreichische Akademie der Wissenschaften, bram.alefs@oeaw.ac.at

Abstract Phoneme boundaries are defined by spectro-temporal properties of the continuous speech signal. In order to detect boundaries automatically, an algorithm is designed that provides segments based on the perceptual difference of succeeding speech periods. The periodicity is estimated by an adapted version of Medan's pitch detection algorithm. The perceptual difference of two periods is approached by the cross correlation of their auditory excitation vectors, which are estimated by a set of 13 ROEX filter shapes. Automatically detected boundaries are compared to manually transcribed speech data for phonemes of the Kiel Corpus.

I. Introduction Speech segments, such as phrases, words and phonemes, consist of one or more acoustic and auditory distinctive segments. The smallest linguistically distinctive segments are phonemes, typical length T=[10ms,>100ms], while the smallest acoustic segments are periods for voiced speech, typical length T=[<4ms, 20ms], and noise bursts for unvoiced speech T<5ms. This study aims to find acoustic segments in continuous speech that can be interpreted as phonemes. An algorithm is designed to detect signal changes that are most apparent to the human auditory system. Audible changes in the acoustical signal are believed to underlie phonetic speech segmentation.

The algorithm is based on the periodicity of the speech signal. Voiced speech is pulsated in nature, i.e. the temporal energy envelope shows a maximal value each period, associated with the glottal pulse. Voiced speech can be described as a sequence of filtered pulses, of which the pulse distance and the filter poles may change from pulse to pulse. In general, unvoiced speech does not have such periodicity and the temporal envelope does not indicate periodicity. The algorithm aims to describe the boundaries of voiced speech, without excluding unvoiced segments a priori.

Considering the speech period as a stationary segment, it can be represented by a series of pure tones, with amplitudes and phases such as obtained by discrete Fourier analysis. Amplitudes are believed to carry the perceptually relevant information, such as formants. Phases contribute to the shape of the temporal envelope. At high fundamental frequencies ($F0$>250Hz) the period may become too short to apply useful spectral analysis. It makes sense to assume a segment of two periods as stationary and to define periodicity at half of the fundamental frequency. The perceptual relevance of a stationary segment can be expressed as the RMS-output of a set of auditory filters.

A pitch detection algorithm (PDA) that takes into account the temporal signal shape was designed by Medan et al. in 1991 [3]. Two succeeding segments S_t and S_{t+T} are considered to be periodic if they are an amplitude modulated version of each other:

$$S_{t+T} = a\,S_t + E \qquad \textit{eq.1}$$

where, a the amplitude ratio and E the estimation error. At position t most likely period T is found at the minimal error E. It can be shown that the minimal error is found at the maximal cross correlation between the segments. Periodicity is detected by an adaptive threshold for the cross correlation (typical: C≥0.8). In order to distinguish the fundamental periodicity from formant periodicity, Medan's PDA evaluates the cross correlation of overlapping segments with maximal period length T_{max}, which are shifted by the proposed period T.

Medan's PDA was designed to estimate speech periodicity with sub-sample accuracy ($\Delta T < 1/f_s$), rather than to detect perceptually relevant changes in the speech signal. In this study an extension is proposed, that focuses on spectral information that is resolved by the auditory system, without assuming local stationarity of the signal a priori. The cross correlation is estimated over auditory excitation vectors, resulting in the most likely sequence of speech periods.

In order to detect segment boundaries, full pitch tracking is not required. In practice, the found period with maximal cross- correlation may not coincide with the fundamental period. As discussed above, for very short periods this may not be desired. The maximal cross correlation of two succeeding segments, within limits of fundamental periodicity, provides an upper limit for the cross correlation of the real periods. The algorithm can be extended for full pitch tracking in further stages, by dividing multiples and removing non-multiple segment lengths.

If the sequence of most likely speech periods is known, speech segments can be derived from similar periods, that show high cross correlation. The boundary of the segment is characterized as a local minimum in the sequence of cross correlations (voiced-voiced boundary) or as the ending of the sequence of cross correlations (voiced-unvoiced or voiced-silence boundary). Within unvoiced speech parts, the sequence of most similar segment helps to find speech segments also. The similarity can be understood as the continuity of the auditory excitation pattern, within a time frame that is sufficiently long to provide such a pattern. Tone bursts shorter than the minimal analysis frame are smoothed to the length of the frame. The algorithm does not correct for this effect yet.

Given the speech segments of most similar periods, phoneme boundaries can be estimated at any desired level. In this study boundaries are set such that at least all phonetic boundaries are represented. Non-stationary phonemes, such as diphtongues are represented in more speech segments, depending on threshold conditions. Minimal phoneme length is set to 10ms for unvoiced speech and silence and 20ms for voiced speech. The automatically detected boundaries are compared to manually positioned boundaries for data of a male and a female speaker.

II. Algorithm The algorithm consists of three stages: peak tracking, period tracking, and phoneme tracking. The peak tracking is introduced to reduce calculation efforts. The phoneme tracking selects speech segment boundaries, when the correlation exceed an adaptive threshold, and eliminates segments smaller than the minimal length.

The period tracking is adapted from Medan's PDA. Succeeding speech segments are windowed by two Hanning windows that overlap at half of their length (i.e. 9% of the integral). The effective overlap is minimized by choosing the center of the window at the maximum of the temporal signal envelope. For each pair of windowed segments the spectrum is estimated by discrete Fourier transform of equal length. Taken into account the spectral spread due to the window, the fundamental is resolved ($F0/2 \le \Delta F < F0$). The perceptual relevance of each spectrum is approached by a set of 13 auditory filter shapes with center frequencies logarithmically spread between f_c=250Hz and f_c=4kHz. The shapes of the auditory filters are described by ROEX (rounded exponential) functions as proposed by Patterson in 1976 [4], with bandwidths according to the ERB-scale. The filters overlap at about their 3dB points, and the outputs of adjacent filters is considered to be independent. Results were not improved by using more sophisticated filter shapes, different center frequencies, or a higher filter density.

Each vector of filter outputs represents the auditory excitation caused by the proposed period. Two vectors derived from succeeding windows, are considered as amplitude modulation versions of each other according to eq.(1). Two segments are considered most similar if the error is minimal, i.e. the cross correlation between the filter output vectors is maximal. For one position t all possible positions $t+T$, are evaluated by increasing the lengths of the succeeding frames. The frames that show maximal cross correlation represent the real period, or at least an upper limit of the cross correlation between two real periods.

The fundamental period is limited to $T_{max}=20ms$ ($F0 \ge 50Hz$), and the minimal analysis length is set to $T_{min}=4ms$ ($\Delta F < 250Hz$). For periodicity with length smaller than T_{min} only multiples of the fundamental period are analyzed. Since the frequency amplitudes of the segments are considered only, the analysis is relative insensitive to small deviations in time domain. In contrast to Medan's pitch

estimation, sub-sample accuracy has limited effect on the overall result. Calculation efforts are reduced by considering only the peaks in the time signal. The peak tracking stage performs a basic peak search to provide these maximums.

Peak tracking The signal is configured such that the glottal pulse coincides with positive peak, and the overall average signal level is set to –40dB relative to the maximal amplitude. For the entire signal all positive peaks within a distance of at least T_{min} are selected, that exceed the average signal level (within a rectangular frame of length $2T_{min}$ centered at the peak). The obtained peak sequence describes all glottal pulses (separated with at least T_{min}) and some low frequency formants. It also includes sufficient noise burst to track fragments in unvoiced speech parts.

Period tracking Each peak is evaluated for it's periodicity with all peaks that follow within the distance T_{max} (i.e. maximal 5). The first window is centered at the peak and it has a length such that it's value is zero at the next peak. The second window has the same length but is centered at the second peak and it's value is zero at the first peak. Since the peaks are maximally separated, the effective overlap between the two segments is minimized in an energetic sense. For both segments spectral amplitudes are estimated by a discrete Fourier transform, with an equal number of frequency bins. The amplitude spectra are filtered with the set of auditory filter shapes, and the filter output vectors A and B are cross correlated such that:

$C = A*B \;/\; |A|\,|B|$ ***(eq. 2)***, where C is the cross correlation for the proposed periodicity. For each peak following, the second window is shifted, and both windows are enlarged. The cross correlation of the filter outputs is estimated subsequently. From all peaks, the one with maximal cross correlation is the most likely candidate for the period. Since the signal is expected to change more within a multiple of periods, the maximal cross correlation is expected at succession of two single periods. The local spectral properties of a formant pulse generally differ from the glottal pulses. Although the maximum does not necessarily represent the real periodicity, it is in all cases the upper limit for the cross correlation between the two real periods.

In order to concentrate on pulses with higher energy it is convenient to start at the highest peak of the signal and to track forward and backward until the maximal cross correlation exceeds the minimal level (threshold: C_{min}=0.8). This may occur at unvoiced speech parts and silences. The period tracking is repeated recursively for the first and last part, until all selected peaks are analyzed.

Phoneme tracking The period tracking stage provides a sequence of maximally correlating speech segments, which, in most cases, coincide with the set of fundamental periods of voiced speech. The recursive period tracking stage divides longer speech signals in several sequences, related to the voiced and unvoiced speech parts with slightly overlapping boundaries. Apart from voiced-unvoiced boundaries, less apparent segments boundaries, such as voiced-voiced changes are expected. It has to be emphasized that phoneme identification may depend on the context also. The task to identify (sub) phonetic segments is left to modern speech recognition technology.

The segment tracking provides a set of acoustically and perceptually distinct elements, that underlay the phonetic structure. Its aim is to find a minimum of elements that still represent all phonemes. In practice some thresholds have to be set: the minimal segment length is set to T_u=10ms for unvoiced speech, and T_v=20ms for voiced speech. Noise bursts, associated with plosive phonemes (*t, p*) are not represented correctly. The boundaries of phonemes are determined by an adaptive threshold for the cross correlation of the auditory excitation vectors.

For a sequence of cross correlation the maximums C_{max} are estimated, with a threshold such that $C_{max} \geq 0.98$, within a distance of at least T_v. Each maximum in the cross correlation sequence represents a phoneme, if the minimum of the sequence is sufficiently low. The threshold for the minimal cross correlation is set to the minimum of C_{max} minus 2.5 times the standard deviation of all following cross correlations, including the next maximum. The position of the segment boundary is set to the period that shows minimal cross correlation to it's successor. The minimal (sub) phonetic length is corrected for overlapping of two independent sequences of correlation (voiced), and for small segments (unvoiced). For each pair of boundaries separated less than T_u, the one is chosen at which the signal has maximal amplitude.

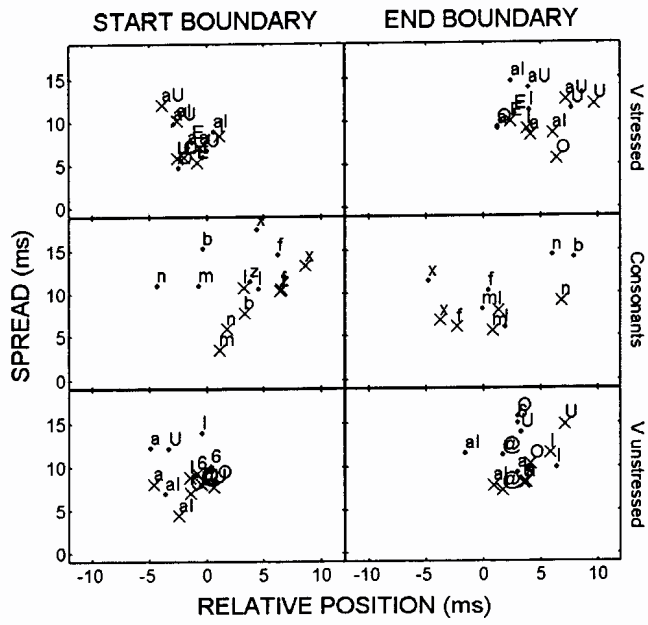

Figure 1. *Position and spread of calculated phoneme boundaries relative to boundaries of manual transcribed speech, for data of a male(×) and a female (•) speaker. Upper: stressed vowels and diphtongues (a, E, I, O, U , aI aU). Middle: voiced and unvoiced consonants (b, f, l, m, n, x, z). Lower: unstressed vowels and diphtongues (6, @, a, I, O, U, aI). Data: read German speech from the Kiel Corpus, speaker K61 and K62.*

III. Comparison The segments found by the algorithm are compared to manually estimated phonemes for a corpus of read German speech (Kiel corpus, CD1), for a male and a female speaker (K61, K62). The corpus is transcribed canonically, such that the boundaries are mostly set at the last zero-crossing before the first voiced pulse of the segment. Phonemes are primarily linguistic entities and the boundaries are not always accurately positioned. The test is designed to compare the algorithm output to the linguistic data, rather than to judge the algorithm's performance. For each of a large set of phonemes, the difference in the position of the manually transcribed start or end boundary to the nearest detected boundary was estimated (P_{START}, P_{END}). Deviations caused by the algorithm are to be found in an bias in the average value, see figure 1. The average relative position (x), is biased such that vowels are slightly enlarged ($P_{START}<0$, $P_{END}>0$). The distributions of the relative positions are approximately normal. The distributions are approximately normal spread and the standard deviation (y) indicates the uncertainty due to inconsequent manual transcription. Results are similar for both speakers.

	N *male*	P START	P END	N *female*	P START	P END
V.stress.	293	-0.7±7.9	5.6± 9.7	294	-0.1± 7.5	3.3±12
Cons.	225	4.4±9.0	1.3±11	213	2.0±13	3.1±13
V.unstr.	341	-1.6±7.6	4.0± 9.7	308	-1.7±10	2.6±12

Table 1. *For each category are shown: the number of phonemes and the difference in milliseconds between the automatic estimated boundary positions and the manual estimated boundaries (*P_{START}, P_{END}*). For each pair* (x±y), *denotes* x *the bias in the average and* y *the standard deviation. The end-positions are biased in all conditions, while start-positions of consonants are positive biased (delayed), and start-positions of unstressed vowels are negative biased (advanced). The standard deviation indicates large uncertainty in the manual transcribed data. See also figure 1.*

References

1. Alefs, B.G., Deutsch, W.A. "Auditory time-frequency analyses applied to phoneme segmentation" FA für Akustik, Österreichische Physikalische Gesellschaft, 2000.
2. Fant, G. "Analysis and synthesis of speech processen" Bertil Malmberg Ed., North-Holland Publishing Co. Amsterdam 1974.
3. Medan, Y., Yair, E., Chazon, D. "Super fine pitch determination of speech signals" IEEE T-SP. 39(1): 40-48, 1991.
4. Patterson, R.D. "Auditory filter shapes derived with noise stimuli" J. Acoust. Soc. Am. 59(3), 640-54, 1976.
5. Institut fur Phonetic und digitale Sprachverarbeitung "Arbeitsberichte Nr 28" AIPUK 28, Universität Kiel, 1994.