Automatic Segmentation of the Speech Signal into Phone-length Elements

W. Jassem
Poznań, Poland

For the purposes of automatic speech recognition, the quasi-continuous signal has to be split up into fragments correlated with linguistic units such as words, morphs or phones. Various considerations of intended applications, technological feasibility as well as the theoretical approach decide what these elements are and how they are discovered, in particular systems. Phoneticians, linguists and acousticians are divided on the subject of segmentability of the speech signal into succesive phonetic entities. Studdert-Kennedy (1981) and Hammarberg (1982), for instance, maintain that phonetic segmentation is only performed at the level of perception and human recognition and that the acoustic speech signal is not segmentable. Such assertion ignores the unshakeable evidence produced, e.g. by Fant (1964) and Reddy (1967) showing that both the original waveform and its transformation into a dynamic spectrogram display unmistakeable points, or brief moments, along the time axis, at which some quite definite changes take place and that these points, or moments, correlate almost perfectly with boundaries between phone-related segments.

A system has been developed in the Acoustic Phonetics Research Unit which consists of a bank of 63 analogue band-pass filters, an interface including an A-to-D converter, and a very primitive 8-K-byte, 8-bit-word minicomputer. This system enables four different kinds of digital spectrograms to be made. For each spectrogram, the signal level is averaged in 4 contiguous frequency bands with a 3-band overlap the effective bandwidth being 320 Hz up to 3560 Hz and progressively larger up to 8310 Hz, resulting in 60 analysis channels. The width of the non-overlapping time windows is 23 ms. (1) The basic spectrogram indicates the signal level in each time-frequency cell above a pre-selected threshold in units of 0.6 dB. (2) The differential spectrogram shows the difference, in positive or negative numbers, between the levels in successive cells in each channel. (3) The difference sign spectrogram only indicates the sign of the difference calculated for (2). Finally, (4), a binary spectrogram signals whether or not the level in each cell exceeds a dynamically varied threshold level. Technical details of the method of obtaining binary spectrograms with the MERA-303 minicomputer are obtained in Kubzdela (1980). The segmentation into phone-length elements here proposed is based on type-3 spectrograms, an example of which appears in Fig. 1. A light sign stands for a minus and a heavy sign for a plus, while unmarked dot indicates defaults corresponding to below-threshold cells. Note that some of the columns in Fig. 1 contain only minusses or only plusses, e.g. column 016, 034, 036, others contain only minusses in the lower frequencies and only plusses in the higher frequences, or vice versa, e.g. 046, 048, while still others are irregular, e.g. 012, 060, etc. in that they consist of more than two same-sign vertical sequences. Basically, segmental boundaries are assumed to occur in columns with one same-sign vertical sequence and, in specific cases, in columns with two same-sign sequences. The formal definitions of the boundaries are as follows:

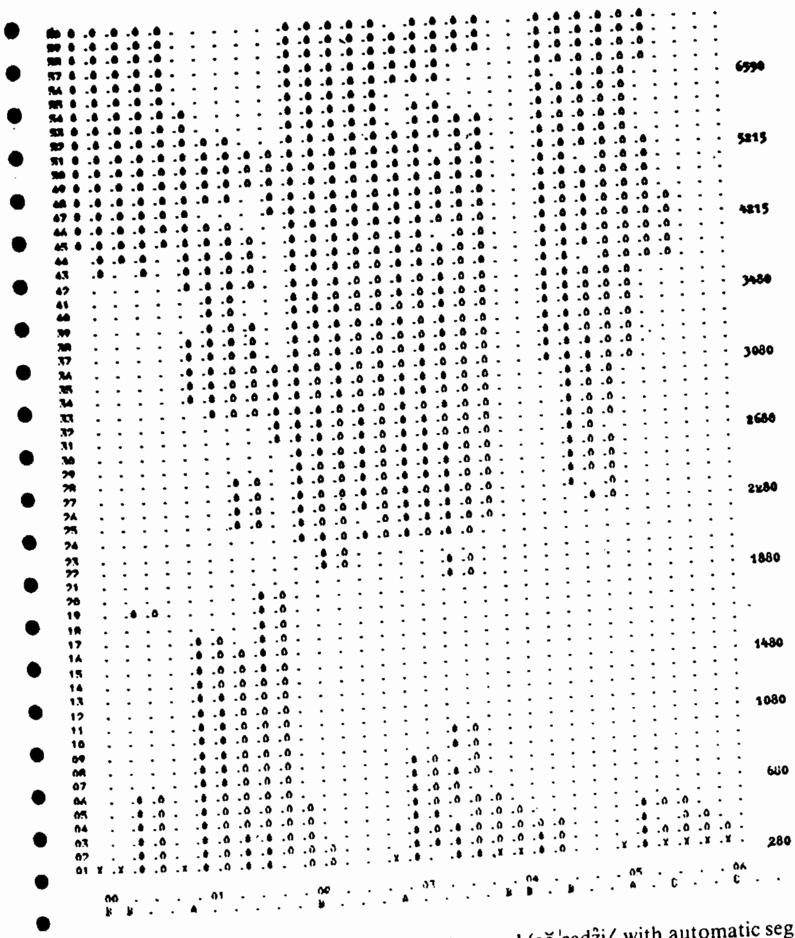

Figure 1. A differential sign spectrogram of the word (sŏ'çedẑi/ with automatic segmentation.

We denote c(k) -- number of same-sign vertical sequences in the k-th column; z(k) -- sign of the highest-frequency vertical same-sign sequence; z(k) equals 0 for a negative sequence and 1 for a positive sequence.

An A boundary appears in the k-th column if

$c(k) = 1_\wedge\{c(k+1) \neq 1_\vee[c(k+1) = 1_\wedge z(k+1) \neq z(k)]\}.$

A BB boundary appears in the k-th and the (k+1)-th column if
$c(k) = c(k+1) = 1 \wedge z(k) = z(k+1) \wedge \{c(k+2) \neq 1 \vee [c(k+2) = 1 \wedge z(k+2) \neq z(k)]\}$.

A C boundary appears in the k-th and the k+ n-1)th column if
$c(k) = c(k+i) \wedge z(k) = z(k+i) \wedge \{c(k+n) \neq 1 \vee [c(k+n) = 1 \wedge z(k) \neq z(k+n)]\}$, with $i = 1,2,...n-1$ and $n \geq -3$.

A D boundary appears in the k-th column if
$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+2) \neq 2$.

An F boundary appears in the k-th column if
$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) \neq z(k+n) \wedge c(k+n) = 2$, with $i = 0,1,2..., n-1$, and $n \geq 1$.

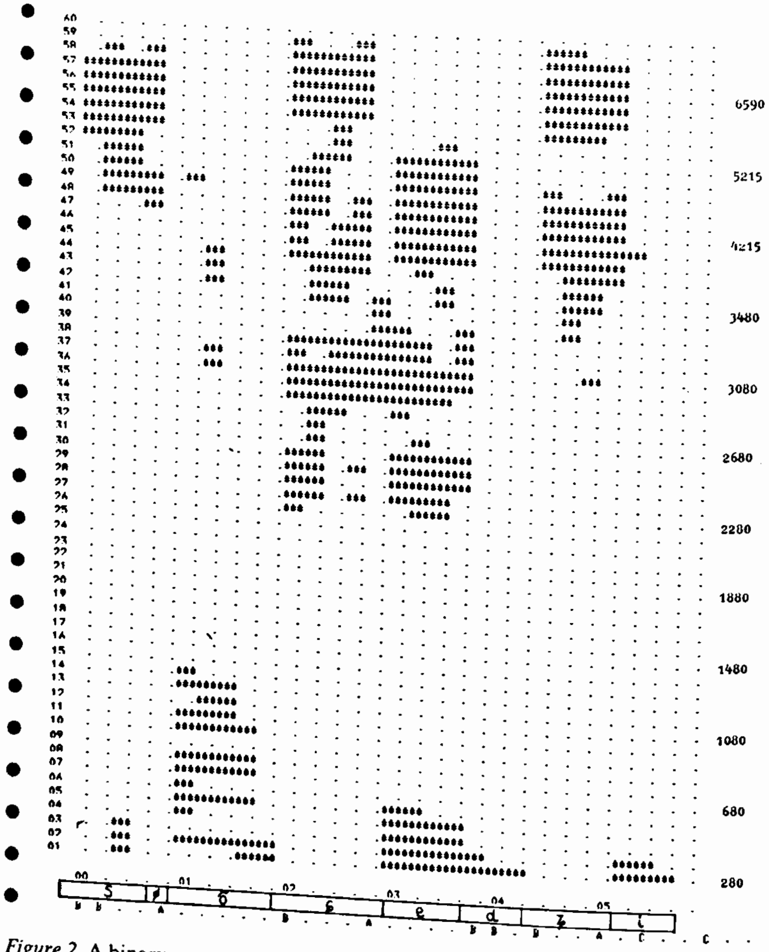

Figure 2. A binary spectrogram of the word /sŏ çedżi/ with visual and automatic segmentation.

A G boundary appears in the k-th column if
$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2$, with $i = C, 1, ..., n-1$, and $n \geq 2$.

The detailed algorithm of segmentation is presented in Jassem, Kubzdela and Domagata (in press). A formal test of the algorithm was performed with one male and one female voice pronouncing, in isolation, 17 Polish words containing sequences of phone types known to be particularly difficult to segment. Overall results were as follows:

	Correct	Misses	False alarms
Male voice	129	25	13
Female voice	122	23	9

Fig. 2 represents a binary spectrogram with visual and automatic segmentation.

The results obtained here compare favorably with those obtained in other ASR systems, particularly in view to its extreme simplicity which enables it to be implemented in a microprocessor. Meanwhile improvements of our system are in progress.

References

Fant, G. (1964). Phonetics and speech research. In: *Research Potentials in Voice Physiology*. New York, 159-239.

Hammarberg, R. (1982). On re-defining co-articulation. *J. Phonetics* **10**, 123-137.

Jassem, W., Kubzdela, H., and Domagata, P. (in press). Automatic acoustic-phonetic segmentation of the speech signal. In: *From Sounds to Words*, Umeå Studies in the Humanities.

Kubzdela, H. (1980). A method of word recognition based on binary spectrograms (in Polish). *IFTP Reports* 15/80.

Reddy, D.R. (1967). Computer recognition of continuous speech. *Journ. Acoust. Soc. Am.* **41(5)**, 1295-1300.

Studdert-Kennedy, M. (1981). Perceiving phonetic segments. In: *The Cognitive Representation of Speech*. T. Myers et al. (eds.), North-Holland Publ. Co., Amsterdam, 3-10.