PHONETIC INTERPRETATION OF ACOUSTIC SPEECH SEGMENTS

Knut Kvale

Telenor Research,
N-2007 Kjeller, Norway
E-mail: knut.kvale@tf.telenor.no

ABSTRACT

A crucial problem in automatic speech recognition is the transformation from the continuously varying speech signal to a set of discrete and abstract phonological symbols. The key question is how much phonetic information can be extracted from the speech signal alone without using prosodic, syntactic, semantic or pragmatic knowledge. To address this issue we have analyzed the phonetic content in spectrally homogenous acoustic segments which are selected automatically. Although the segmentation algorithm is language independent and needs no training session, we found that the obtained acoustic segments could be given phonetic interpretations.

1. INTRODUCTION

In the last few years impressive progress has been made in *spoken language systems* (SLS) which make it possible for people to interact with computers using speech. The SLS technology integrates techniques of automatic speech recognition (ASR), natural language processing and human interface facilities. A crucial problem for ASR is the transformation from the continuously varying speech signal to a set of discrete and abstract phonological symbols.

Although listeners tend to perceive speech as discrete sounds following each other in temporal order, the mapping between acoustic events and a linguistic representation is complex, non-linear, irreversible and only partly understood. Thus, the discreteness is not signalled by the stimulus but is imposed on that stimulus by a listener.

The design philosophy of many automatic speech recognition systems has therefore been based on the belief that the acoustic signal does not provide sufficient information to identify the linguistic content of an utterance. Thus, prosodic, syntactic, semantic and pragmatic knowledge has to be utilized to recognize an utterance.

By contrast, experiments in speech spectrogram reading, e.g. [1],[2], have demonstrated that phonemes are accompanied by acoustic features that are recognizable directly from the speech signal without additional knowledge sources. This paper pursues this issue further by analyzing automatically derived stable portions of the speech signal.

2. ACOUSTIC SEGMENTATION

A *segment* is a linear unit anchored in a short stretch of speech by a set of relatively unchanging phonetic feature-values [3]. Thus, *segmentation* can be defined as dividing the speech signal into directly succeeding, non-overlapping stable parts.

Algorithms for automatic *acoustic segmentation* rely on the acoustics only, i.e. they do not assume any phonological information. There are many advantages of acoustic segmentation compared to phonemically based segmentation. Firstly, the speech segments are characterized by acoustic, language independent properties, which can be derived automatically. That is, the calculations are entirely based on signal processing and hence there is no need for explicit modelling or any prior phonological knowledge of the language. Secondly, the automatic subword generation is deterministic in that identical waveforms will be segmented into the same acoustic subword. Thirdly, the acoustic segments often contain highly correlated frames and can hence be quantised, i.e. represented by less data, without losing essential information.

In this paper we analyze the acoustic segmentation calculated by the *Constrained Clustering Vector Quantization* (CCVQ) algorithm [4],[5]. This algorithm recursively computes all possible segment combinations and represents each segment by its centroid, (i.e. its mean spectrum with the present distortion measure). The optimal segment sequence minimises the differences between the spectral frame vectors and the centroid within each segment. That is, the consecutive acoustic segments which yield minimal overall intra-segmental distortions are found. The obtained segments thus exhibit the maximal acoustic homogeneity within their boundaries and the frames within a segment are highly correlated, i.e. steady segments are located.

Phonemically defined units may contain many spectrally homogenous or quasi-stable areas. Thus, acoustic segmentation algorithms may often provide an *oversegmentation* (o.s), i.e. more segments than phonemic labels. As an example, *figure 1* displays a speech waveform and the corresponding broad-band spectrogram which is automatically segmented with the CCVQ-algorithm with 100% o.s., i.e. the number of acoustic segments is forced to be twice the number of phonemes in the utterance. The speech signal in figure 1 is manually segmented and labelled with SAMPA-symbols [6] according to the conventions described in [4],[7].

3. QUALITATIVE EVALUATION

The qualitative analyses of the CCVQ-algorithm were carried out on the Norwegian EUROM0 recording [4],[7]. With 100% oversegmentation typical general trends were (see [4] for details):

- *Plosives* were most often segmented into a closure part and a burst part, such as /k/ in /O:kek/ in figure 1. However, when voiceless plosives succeeded an /s/, as /sp/ and /sk/ in figure 1, the plosive release was weakened and was not marked as a separate segment. If the closure contained some voicing, this was also separated as one segment. Often some alternatives for the beginning of the closure and the end of the burst were given. If the plosive release contained both a burst and an aspiration part, these were marked as two separate segments.
- *Vowels* realised with an amplitude that increased evenly to a maximum value and then decreased towards the next phoneme often contained formant-transitions which were detected by the acoustic segmentation and an acoustic segment boundary was placed near the amplitude top as in the first /i/ in figure 1. (Marking the "centre" of the phonemes is useful for e.g. consistent diphone segmentation for text-to-speech synthesis [8]).
- In the transition from *vowel to silence* the acoustic segmentation algorithm calculated two or three boundaries as for /O:k/ and /ek/ in figure 1. The first one was placed where the intensity reduction began in the higher frequencies, the second (optional) one was placed where almost no intensity was registered in the spectrogram, and the third one was placed where no intensity at all was detected in the spectrogram.
- Segments containing *extralinguistics* (e.g. creaky voice, epenthetic silence, epenthetic sound and lipsmack) were

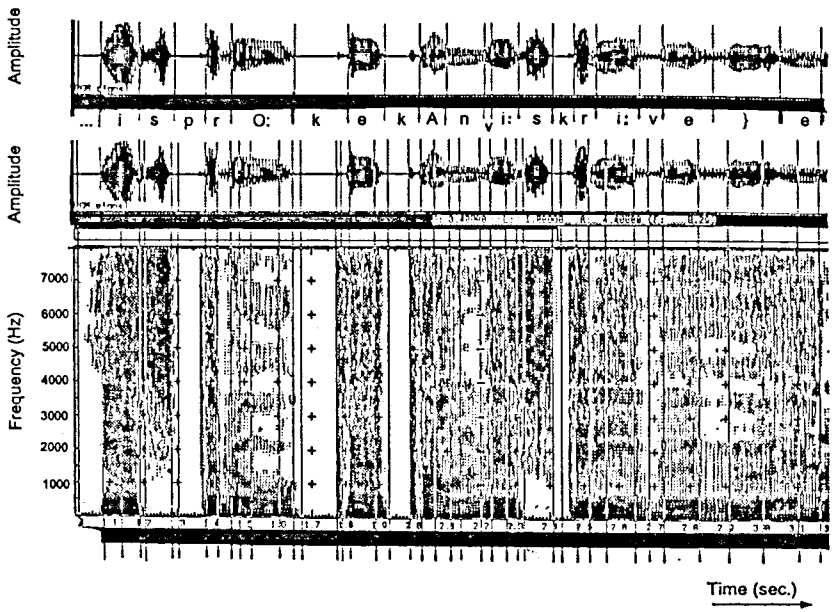

Figure 1 *The sentence "i språket kan vi skrive ue(ndelig)" (=in language we can write inf(initely)) is manually segmented and labelled with SAMPA symbols [6] shown under the waveform in the top row of the figure. In the waveform below and in the broad band spectrogram the acoustic segmentation boundaries with 100% over-segmentation are shown. After [4].*

marked off separately. In figure 1 we notice that the creaky voice area between /e/ and /}/ is one acoustic segment.

• When the apico alveolar tap realisation of /ɾ/ showed up in the spectrogram with an extra voiced sound with formant structure [9], i.e. an epenthetic schwa as in /sprO:/ and /skri:v/ in figure 1, the schwa and [r]-closure were segmented into separate acoustic segments.

4. OVERSEGMENTATION

Obviously it is preferable to keep the oversegmentation factor (o.s.-factor) as low as possible while still achieving high coincidence with manual segmentation. This section summarizes the performance of CCVQ-segmentation as a function of oversegmentation:

• Boundaries computed with a lower o.s.-factor remained fixed when increasing the o.s.-factor. That is, the effect of increasing the o.s.-factor was to split the segment(s) with highest intra-segmental distortion. Actually, spectrally stable segments were not divided even with 200% oversegmentation.

• The CCVQ-algorithm searched for stable segments, and the boundaries were placed in transient areas because vectors from these areas increase the intra-segmental distortion. As the o.s.-factor increased, transition areas could be segmented into several short acoustic segments, providing several alternative boundaries. This reflects the segmentation problem of placing a boundary between two sounds at one, single, "correct" time instant.

• With more than 75% o.s., the acoustic segmentation obtained high coincidence with the corresponding manually placed boundaries. The few deviations from manual segmentation were mainly due to:

i) Some *half-way*, *mid-point*, or *symmetric* conventions used in the manual segmentation, e.g. the convention in [4] of placing the boundary in the middle of the creaky voice area between two vowels (instead of at the end of the segment where abrupt changes often occur). If this area was spectrally stable, the acoustic segmentation assigned boundaries at the ends of it.

ii) "*Impossible cases*", where no boundary cue was seen in the waveform or spectrogram, and the human labeller has placed the boundary rather arbitrarily or based the decision on listening only.

iii) "*Squeezed in segments*", i.e. a phoneme which is perceived when listening to it in context but which is without any corresponding visible acoustic cues in the waveform or spectrogram, was often squeezed in as a very short segment between the phonemes with clear acoustic cues, e.g. as /v/ in /nvi:/ in figure 1.

5. CONCLUSIONS

The CCVQ-algorithm isolated spectrally stable portions of the speech signal. The stable segments were not divided even with a high degree of oversegmentation.

When the number of acoustic segments was forced to be twice the number of phonemes in an utterance, most of the acoustic segments obtained by the CCVQ-algorithm could be given a phonetic interpretation. In addition, quantitative analyses in [4] have showed that the acoustic segment boundaries coincided equally well with the corresponding manual segmentation for English, Danish, Norwegian and Italian (manually annotated by native phoneticians).

Since the acoustic segmentation algorithm is capable of isolating identifiable sub-phonemic segments consistently, it can be useful for speech analysis and automatic speech recognition based on acoustic subwords. The CCVQ-algorithm may also be used as a language independent pre-segmenter tool for manual segmentation of e.g. diphones for text-to-speech synthesis. When this tool is accompanied by conventions for which boundaries to select for the various phoneme transitions, it will reduce the randomness in manual segmentation.

REFERENCES

[1] Zue, V.W. and Cole, R.A. (1979), "Experiments on spectrogram reading", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 116-119.

[2] Zue, V.W. (1989), *Speech Spectrogram Reading - An Acoustic Study of English Words and Sentences*, Course at University of Edinburgh.

[3] Laver, J. (1994), *Principles of phonetics*, Cambridge University Press.

[4] Kvale, K. (1993), *Segmentation and Labelling of Speech*, Doctoral thesis, Norwegian Institute of Technology.

[5] Svendsen, T. and Soong, F.K. (1987), "On the Automatic Segmentation of Speech", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 3.4.1-4.

[6] Wells, J.C., et al (1992), "Standard Computer-Compatible Transcription", in *ESPRIT PROJECT 2589 (SAM): Final Report; Year Three; 1.3.91-28.2.92*, SAM-UCL-037.

[7] Kvale, K. and Foldvik, A.K. (1991), "Manual Segmentation and Labelling of Continuous Speech", ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 37.1-5.

[8] Kvale, K. (1995), *Manual segmentation of logatomes for diphone-based text-to-speech synthesis*, Scientific Report 2/95, Telenor Research.

[9] Kvale, K. and Foldvik, A.K. (1992), "The multifarious r-sound", Proc. International Conference on Spoken Language Processing, pp. 1259-1262.