TOWARDS A VOICELESS SPEECH SKETCH

N. R. Kew

University of Sheffield, Department of Computer Science, Sheffield, England.

Abstract

The SYLK project aims to combine knowledge-based and statistical approaches in the task of Automatic Speech Recognition (ASR). The two approaches are integrated via an entity known as the SYLK "Refinement Test", which interprets arbitrary, and often knowledge-based low-level processing in a statistical framework. In this paper, we first introduce the refinement test. We then proceed to develop a class of tests aimed at the task of plosive discrimination, and substantially based on the notion of a Voiceless Speech Sketch.

1. Introduction

The SYLK project is a new approach to ASR, addressing the problem of the acoustic-phonetic transformation. An overview of SYLK is given in Green et al [1] in this volume. In this paper we consider the lowest level of SYLK, which includes signal processing, semi-symbolic description (the Speech Sketch) and feature extraction, up to the point where low-level descriptions are statistically trained.

In Section 2, we first formally define the Refinement Test. We follow this with some explanatory discussion. Two contrasting examples, described in subsection 2.3 may prove for many readers the most informative part of this section.

In Section 3, we develop a Voiceless Speech Sketch. We describe how it may be used to construct refinement tests in the task of plosive discrimination. We concentrate particularly on a frequency-domain sketch, the Burst Profile, derived from the short-time Fourier transform of a burst waveform.

2. The Refinement Test.

2.1. Definitions

2.1.1. A FEATURE VECTOR SPACE is a quadruple (V, r, n, i),
where
V is a metric space.
r (the RANK) is the order of elements of V.
i is an imbedding of V into a finite-dimensional Real metric space. It is constrained to preserve the metric on V.
n (the DIMENSION) is the dimension of the image i(V) (usually equal to the rank).

We note that the requirement to imbed in finite-dimensional real space imposes a constraint on V.

2.1.2. A TRAINING on a feature vector space **S** is a quintuple **(N, H, F, k, W)**,
where:
D (the DOMAIN) is a set, which may be a structured set (as below) of phonetic/phonemic classes.
H (the HYPOTHESES) is a mutually exclusive and exhaustive set of subsets of **D** (specifically the classes between which the training discriminates).
k (the ORDER) is the cardinality of **H**.
F (the TRAINING) is a set of **k** probability density functions (pdf's) on **S**, one representing each Refinement. These pdf's may be represented either parametrically using appropriate Normal and/or discrete models, or nonparametrically using kernel and nearest neighbour estimation techniques [2].
W is a Speech Database (discussion of which is beyond the scope of this paper).

This definition describes the training in the full generality allowed by Boucher's framework [2]. However, SYLK is based on a syllable model (described in [1] and [3]) which represents a structure on syllabic component classes, and a structured training is therefore appropriate. We define:

2.1.3. A SYLLABIC TRAINING is a training in which **D** represents a node of the SYLK syllable model, and **H** comprises nodes (or sets of nodes) below **D**. **H** may then also be regarded as representing the Refinements of **D** in the syllable model.

2.1.4. A TEST is a sextuple **(D, R, P, V, C, F)**,
where
D (the DOMAIN) is the domain of the Training.
R (the REPRESENTATION) is a representation of a fragment of speech signal (eg Waveform, parameter trace, spectrogram, short-time Fourier transform, formant track). It is constrained by an assumption that it incorporates exactly one instance of a speech event appropriate to the Domain.
P (the PROCESS) is a function **P:R->V** computing a feature vector.
V (the DESRIPTION SPACE) is a feature vector space.
C (the CONTEXT) is a feature vector space.
F (the TRAINING) is a training on **VxC**.

2.1.5. A CONTEXT-INDEPENDENT TEST is a test in which the context is null.

We observe that any test **T** may be represented as a context-independent test **T'** with **R' = RxC, P' = PxIdentity**, and **V' = VxC**.

2.2. Discussion

We note that the constraint on the Representation often requires the possibility of a null event, which is indeed defined at appropriate nodes of the syllable model. We see too that it clearly requires some preprocessing to derive one. This may be based on the segmentation of the signal performed by SYLK's initial HMM recogniser (when the segments may be treated as rigid or used more flexibly as a guide to the location of an event), or **R** may be created according to a result from another test. The context (if present) is typically derived from feature vector(s) generated by one or more previous test(s).

More loosely, a test may be regarded as comprising a Process and a Training, and acting at a syllable model Node. In these terms, speech knowledge is built into the creation of a test Process, whilst the training ensures that the updating of the state of belief is (locally) optimal, and represents a major advance when compared to a rule-based framework.

The particular utility of the test is that it provides a well-defined interface for low-level processing which is (a priori) totally arbitrary, in a statistically optimal Bayesian framework. The test developer can concentrate on the tasks of pattern recognition and feature extraction, without reference to the interpretation of his results.

We note that the training of individual tests does not guarantee global optimality unless all tests are orthogonal. This is a particular problem when the same data is re-used in more than one test. A discussion of the treatment of non-orthogonal tests is beyond the scope of this paper

2.3. Examples

2.3.1. The distribution of energy in a plosive release yields articulatory information. An alveolar burst typically has higher energy than a bilabial burst, particularly in the higher frequencies. A test process is therefore defined characterising the energy in the crucial region from 800 to 4000 Hz as a feature vector (mean, variance, gradient). Here a rule-based approach might define rules such as
{(low mean, negative gradient) -> bilabial}, and
{(high mean, positive gradient) -> alveolar, or possible velar}.
In the context of a refinement test, the training provides a precise probabilistic interpretation of the feature vector.
This process is used in several tests, by training at different nodes.

2.3.2. (This test is currently purely hypothetical)
A neural net is trained over burst profiles to discriminate between phonemes /b/, /d/, /g/ using instances in syllable onset position. This creates a test process appropriate to node D (voiced plosive onset) in the syllable model. Here the Description Space is a set comprising just three elements, the neural net outputs. The SYLK training (which bears no relation to the neural net training) requires distances between all these elements through imbedding in 3-space, so this is an example of a test with rank 1 and dimension 3. The SYLK training is closely related to the confusion matrix for the neural net.

3. The Voiceless Speech Sketch and Plosive Discrimination Tests.

A number of refinement tests for plosive discrimination have been developed by an expert system approach, following the detailed study of burst profiles by O'Brien [4]. There are four stages to this process:

(1) Detection and accurate location (in time) of a plosive burst.
(2) Transformation of the signal to a profile (frequency domain representation).
(3) Characterisation of the profile in simple terms (speech sketch).
(4) Extraction of meaningful features in the profile (test processes).

3.1. Time-domain processing (stage 1).

This is the most substantial task, and as space precludes a detailed account of both this and our main theme of frequency-domain processing, our discussion lacks the full depth the problem deserves.

3.1.1. A Parameter Trace.

A parameter trace commonly used by phoneticians to indicate frication is the zero-crossing rate ZC0. Whilst undoubtedly a valuable asset to the human spectrogram reader, its use in ASR poses certain problems:

(1) Where speech is digitised with a non-zero mean (DC component), a weak burst may fail to register in ZC0. This has been observed by Simons in the Edinburgh RM1 Corpus of Phonetically Dense Sentences.
(2) Levels of background noise can give a spurious high ZC0 value in quiet moments. This is clearly in evidence in Timit.

Problem (1) can be overcome by using the extrema rate ZC1 (zero crossings in the first derivative of the signal) in place of ZC0. In principle, ZC1 measures the frequency of the highest frequency component of a signal with little reference to amplitude. However, this seriously exacerbates problem (2), to the extent that ZC1 often decreases from a silence through a burst.

We adopt a new approach to problem (2). Rather than simply count crossings, we measure the amplitude of each crossing (i.e. the absolute difference of the signal values before and after). This masks out silences very effectively, but has the side effect of emphasising voiced speech. Whilst this effect is small, we nevertheless seek to reduce it further by dividing each amplitude by the interval since the previous crossing. This has relatively little effect. We call our new trace ZC*.

Figure 1 shows examples of parameter traces ZC0 and ZC*, for a weak /b/ and a strong /t/. All are smoothed using a Normal window with an SD of 64 samples.

We see that ZC* behaves well in both cases, whilst ZC0 in contrast gives no indication of the weak burst.

3.1.2. Burst Location

(This section summarises briefly a test process)

A burst plosive is characterised by a pause (closure) followed by a sudden sharp peak in ZC*. Thus a test process for a burst consists in describing such events in ZC*. This also offers some discrimination between plosives and imposters such as affricates (/ch/) or consonant clusters (/ts/) for which the peak in ZC* may be more sustained or diffuse. An associated test process Voice Onset Time is also used to detect voicing. The precise location of a burst may best be determined by accurately locating its onset. We achieve this by looking for a peak in the smoothed first derivative of ZC* (computed by convoluting ZC* with the first derivative of a Normal distribution). We employ a multiscale process, in which the burst onset is located approximately in a highly smoothed trace and then tracked to a finer scale. This coarse-to-fine tracking is unusually simple because we can employ the knowledge of the preceding pause and simply track forward without the danger of a spurious peak.

Figures 1 (d),(h) show clearly well-defined burst onsets at the coarse scale.

3.2. Frequency Domain Representation

A frequency domain representation is obtained using an FFT on the burst waveform. This involves some compromise between the need to focus on what may be a short burst, and the need to use sufficient data to see the features described by O'Brien (who recommends a longer time window). At present, we accomplish this using a rectangular window over 256 data values, representing 16 ms at 16kHz sampling rate. Prior to transforming, the data is put through a pre-emphasis filter of 0.95. We find this gives profiles which are visually similar to O'Brien's, and exhibit the salient features she identifies. In view of the frequency-domain processing described below, the side-lobes arising due to a rectangular FFT are not considered a serious drawback.

3.3. A Frequency-Domain Speech Sketch

There are three stages to the processing of a burst profile prior to explicit feature extraction:
(1) The data is convolved with a Normal distribution and its first two derivatives, to obtain smoothed representations of the data and its derivatives.
(2) Each of the resulting curves is parsed for extrema.
(3) The extrema data is combined to give a new piecewise linear representation of the data, on which the parsing information is superimposed. This representation is called the Frequency-Domain Speech Sketch, and is used as the Representation Space for a number of knowledge based refinement tests.
We observe that extrema in the smoothed data and derivatives represent respectively extrema of the data, and of its gradient and curvature. Hence the parsing of these describes precisely the most interesting set of points in a smooth curve.
Figure 2 illustrates this process (but omits parsing information). We see that the data is smoothed whilst retaining its general form and salient features, and that the speech sketch representation reflects closely the smoothed data.

3.4. Tests using the Burst Profiles.

3.4.1. Characterisation of Peaks and Dips.
One further stage of profile parsing is involved in the implementation of these tests. A peak in the profile is described as a vector (x, y, w, h, a), where:
x is the frequency at which the data maximum occurs.
y is the data value at x
w is the frequency difference between the following minimum and the preceding maximum in the first derivative, and is a measure of the "width" of the peak.
h = y - 0.5(y- + y+), where y- and y+ denote the data values at neighbouring maxima in the second derivative - the "base" of the peak. We see that h represents the relative height of a peak above the surrounding data.
a = w*h is a relative measure of the "area" of a peak.
(n.b. all references to "data" in the above should read "smoothed data").
A dip is similarly characterised, with the values of h and a being negative.

3.4.2. Use of Peaks/Dips in Test Processes.
The analysis we have described is now suitable for use in refinement test processes. We illustrate this by means of an example:
A velar plosive burst is often characterised by spectral prominences due to 1/4 and 3/4-wavelength cavity resonances. These appear in the burst profile as peaks in the frequency ranges 1-2 kHz and 3-5 kHz, and at a frequency ratio of approximately 3.0. An example is shown in Figure 3(a). Test process BIMODAL_1 checks for the presence of such peaks, returning TRUE if they are found and one of them is also the maximum data value in the range 0-5kHz, FALSE otherwise.
Unfortunately, peaks in these frequency ranges may be found due to other reasons, including general (random) variations. Figure 3(b) shows a typical alveolar profile, showing small spectral prominences at 1.64 and 2.58 kHz (the subject of another refinement test) and a broad peak around 4.1 kHz. Although this is clearly not bimodal, BIMODAL_1 returns TRUE. A second test process BIMODAL_2 describes the peaks, returning a feature vector comprising measures of their sizes and the ratio of their frequencies. The training for BIMODAL_2 then distinguishes between the genuine bimodal profile (a) and the imposter (b).
We note that the framework of SYLK does not permit the combination of BIMODAL_1/2 in a single test. This is because the training of BIMODAL_2 requires a feature vector (which, by definition, has fixed dimension) which is meaningless in cases where BIMODAL_1 has returned FALSE.

3.5. Further Reading

In this paper, we have described the notion of refinement test and frequency domain speech sketch. Descriptions of several such refinement tests, and preliminary results, may be found in Kew [5].

4. References

[1] PD Green, AJH Simons, PJ Roach "The SYLK Project: Foundations and Overview", this volume.

[2] LA Boucher, PD Green "Syllable Based Hypothesis Refinement in SYLK", this volume.

[3] AJH Simons "Object Oriented Data Structures", SYLK working paper #2, Department of Computer Science, University of Sheffield.

[4] SM O'Brien "The Speech Knowledge Interface: Observations on the identification of plosives", HCC Report HCC/L/41, LUTCHI, University of Loughborough.

[5] NR Kew, PD Green "A Scheme for the Use of Syllabic Knowledge in Statistical Speech Recognition", Proc SST-90.

Figure 1: Time-Domain Traces for Burst Detection

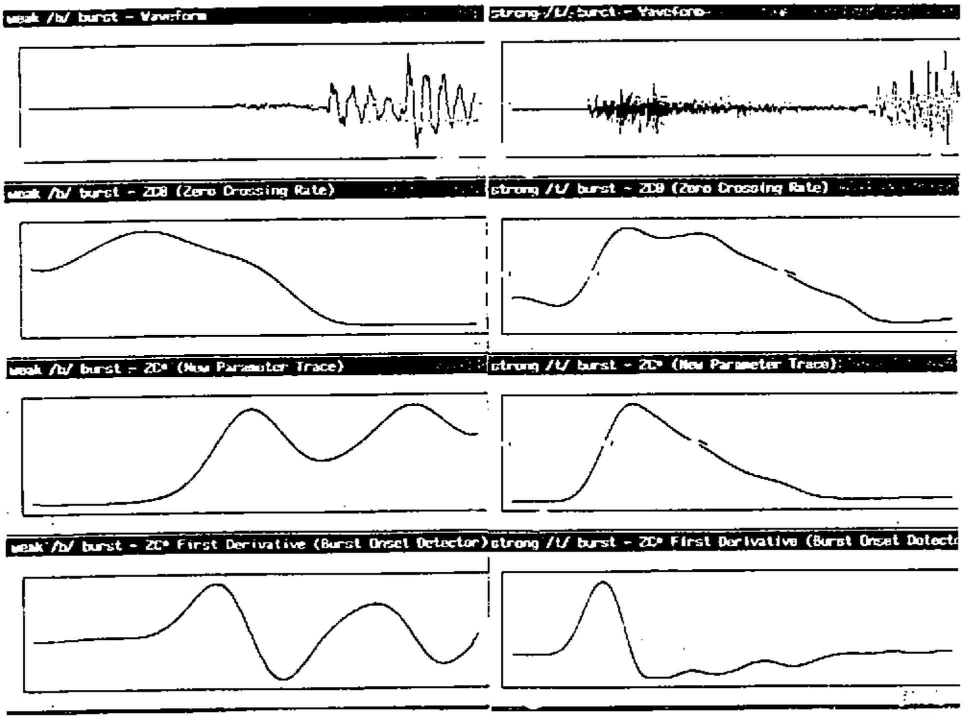


Figure 2: Processing of Burst Profiles

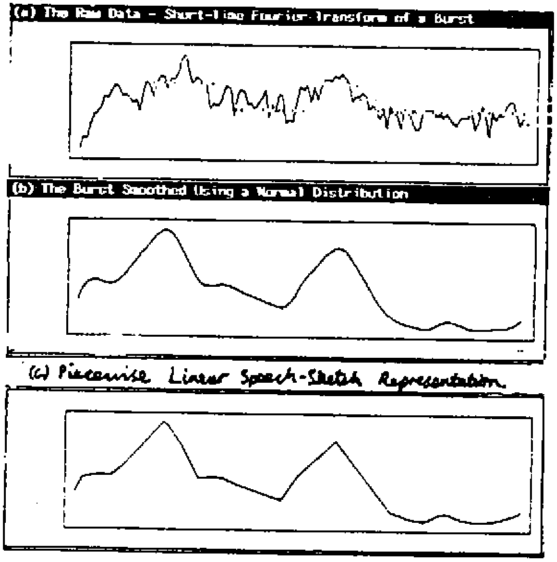
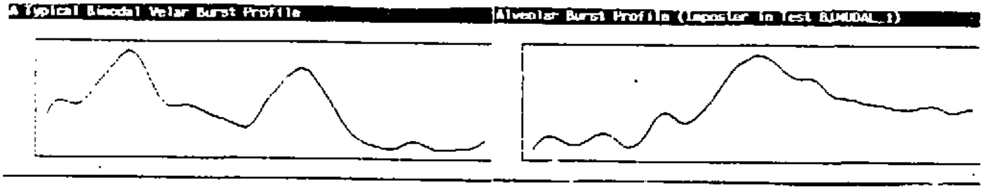


Figure 3: Bimodal Burst Profile and Imposter


A Typical Bimodal Velar Burst Profile
Alveolar Burst Profile (Imposter in Test BIMODAL_1)