Soft Harmonic Masks for Recognising Speech in the Presence of a Competing Speaker

André Coy and Jon Barker

Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, United Kingdom
{a.coy, j.barker}@dcs.shef.ac.uk

Abstract

The paper addresses the problem of recognising speech in the presence of a competing speaker. It uses a two stage ‘Speech Fragment Decoding’ system. The system works by first segmenting a spectro-temporal representation of the mixture into a number of fragments, such that each fragment is dominated by a single source. An ASR search is then extended to find the combination of speech model sequence and fragment subset that best fits a set of clean speech models. This paper extends previous work by combining ‘Speech Fragment Decoding’ with soft missing data techniques to better handle spectro-temporal representations of each source. In that work pitch tracking was used to extract a single fragment for each source. The remaining inharmonic regions are segmented using techniques borrowed from image processing. A shortcoming of this system is that at certain time-frequency points the harmonic energy cannot be reliably labelled as either one source or the other. This typically happens when the pitch of one source is harmonically related to the other - it would also be a problem if the pitches were close in value as they may often be for mixtures of matched-gender speakers. One solution would be to break the harmonic regions into smaller fragments, but this would add to the complexity of the decoding. In the current work we solve the problem by introducing soft decisions in the harmonic fragment regions.

The following section examines the processes of generating the fragments. Section 3 describes the experiments performed using the fragments generated in section 2 and finally, the results and discussion are presented in section 4.

1. Introduction

Recognition of speech in the presence of other sound sources remains a challenging problem. Specific solutions do exist but all impose constraints. Some rely on the presence of multiple microphones [1]. Others assume the noise ‘background’ has temporal dynamics that are very different to those of speech [2]. Another set of techniques attempts to directly model the combined speech plus noise signal, but in order to keep the problem computationally tractable invariably assumes some a priori knowledge of the structure of the noise sources [3]. None of these techniques work well when applied to single microphone speech in general everyday listening conditions. And, they all fail badly in certain ‘pathological’ conditions, such as speech in the presence of other speakers - the so-called ‘cocktail party’ condition [4].

An alternative approach to handling non-stationary noise is ‘speech fragment decoding’ [5]. This approach has been largely motivated by studies of the auditory system and its ability to form the perception of separated sound sources given a mixed acoustic signal [6, 7, 8]. The approach works by segmenting a spectro-temporal representation into a number of spectro-temporal ‘fragments,’ where a fragment is a region dominated by a single environmental source. Each fragment is either dominated by the target source or by one of the background sources, but the labelling is not known. A top-down search is employed to simultaneously solve this foreground/background segmentation and find the best word sequence.

In the previous work [9] the Speech Fragment Decoder (SFD) was employed to recognise speech mixed with another speaker (of opposing gender) at 0 dB. In that work pitch tracking was used to extract a single fragment for the harmonic regions of each source. The remaining inharmonic regions are segmented using techniques borrowed from image processing. A shortcoming of this system is that at certain time-frequency points the harmonic energy cannot be reliably labelled as either one source or the other. This typically happens when the pitch of one source is harmonically related to the other - it would also be a problem if the pitches were close in value as they may often be for mixtures of matched-gender speakers. One solution would be to break the harmonic regions into smaller fragments, but this would add to the complexity of the decoding. In the current work we solve the problem by introducing soft decisions in the harmonic fragment regions.

This work was funded by EPSRC grant GR/R47400/01
and the peaks representing harmonics of the active source(s) are removed. The remaining peaks are chosen as initial candidates for the pitch(es) in that frame.

The lags of the dominant peaks in the autocorrelation function of each channel are measured. If the channel lag matches within 5% of the candidates from the summary then that candidate is chosen as the pitch estimate for that channel; if not, the channel is considered unreliable and no estimate is produced. If there is more than one matching estimate to a single peak from the summary then the one that has the highest percentage match is chosen as dominant in that channel. A ‘1’ is placed in the mask at this spectro-temporal point to indicate a high level of certainty in the dominance of the chosen source. However, when there are two sources active in a frame (indicated by two peaks in the summary) the source with the highest peak is chosen. The decision however, has to be softened because there is much less certainty about which source is actually dominant in the channel. Rather than expressing total confidence in the dominance of the chosen source, a soft decision is made by considering the relative contributions of both sources in that channel. The contribution, $c_i$, of source $i$ is taken to be $p_i \times h_i$, where $h_i$ is the height of the channel’s peak and $p_i$ is a number between 0 and 1 indicating how closely the lag of the channel’s peaks matches the corresponding peak found in the summary. This is done for both sources and the ratio ($x$) of the greater and lesser contributions is taken. Using a sigmoid mapping ($f(x)$) the ratio of contribution values is then compressed to a number between 0 and 1. The sigmoid is centred on 1; the point of maximum uncertainty. The mapping function is given by:

$$f(x) = \frac{1}{1 + e^{-\alpha(x - \beta)}}$$

where $\alpha$ is the slope of the sigmoid and $\beta$ is the centre. The value for $\alpha$ is chosen experimentally. The values derived from this mapping tend toward binary values as $\alpha$ tends to $\infty$. The closer the height of the two peaks, the closer $f(x)$ will be to 0.5, indicating greater uncertainty. By using these soft values in the mask, the contribution of both sources is taken into account and the possibility of pitch estimation errors is acknowledged. An example of an utterance, the soft harmonic mask for a single source and the inharmonic fragments of both sources is shown in figure 2. No uncertainty measurement is attached to the inharmonic fragments.

By focusing on mixtures of male plus female speech the male and female harmonic regions can be estimated directly by clustering the pitch values of the time-frequency elements into a low frequency and a high frequency cluster. The more general problem of dealing with same-gender mixtures where the individual pitch contours are not easily separable is discussed in Section 4.

2.2. Unvoiced Fragments

Although the main focus of this work is the creation and use of soft harmonic masks there is still room to show that there are useful data found in the inharmonic regions of the utterances. The process for inharmonic region segmentation is outlined here (see [9] for a full description).

As a consequence of the time-frequency concentration of speech energy, when multiple sources are present, energy regions of the individual sources only partially overlap. High-energy features of the individual sources often appear as separated peaks of energy which can be segmented using an algorithm employing the watershed transform [11]. The algorithm groups regions of high energy that fall into ‘intensity valleys’.

Further grouping is done by grouping sound elements that share a common onset time, since they are likely to belong to the same source. An example of an utterance, the soft harmonic mask for a single source and the inharmonic fragments of both sources is shown in figure 2. No uncertainty measurement is attached to the inharmonic fragments.

3. Experiments

3.1. Experimental Setup

The corpus of monaurally mixed speech developed for testing the SFD was employed in this task [9]. Digit strings (taken from the clean utterances in the Aurora 2 corpus - test set A) uttered by speakers of differing gender were end-pointed and approximately matched for length. The signals are then artificially mixed at 0 dB in the time domain; this results in 484 utterances which are used for testing.

Whole word gender dependent spectral HMMs were trained
using clean training data. The HMMs had 16 states in a straight-
through topology. Each state was modelled with a mixture of
7 Gaussian distributions with diagonal covariance matrices. A
single state silence model was constructed to model inter-digit
pauses. When testing, either male models or female models are
employed depending on the gender of the target utterance.

An approximate model of auditory nerve firing activity is
used to generate the features used in both training and recogni-
tion. A bank of 64-gamma tone filters (centre frequencies
equally spaced on an ERB scale from 50 Hz to 3850 Hz) were
used to create acoustic vectors. Using a 1st order filter with
an 8 ms time constant, the instantaneous Hilbert envelope is
smoothed and sampled at a frame rate of 10 ms. Cube root
compression was then applied to the envelope values to create
the final representation.

An \textit{a priori} mask is developed by comparing the time fre-
quency representations of the \textit{unmixed} signals. This mask con-
ists of the spectro-temporal regions of the mixture in which
the target is undisturbed by the masker. The results of missing
data recognition performed using these masks are shown in the
first column of the tables (Table 1 - 3). This result represents
the performance expected when perfect segregation of the tar-
get and masker is achieved.

3.2. Recognising Harmonic Regions with Soft Masks

Previously the masks generated for recognition were discrete;
marking regions as belonging to the target or masker with a
strict binary decision. These discrete masks were formed by
putting a 1 in the mask where a spectro-temporal ‘pixel’ is
known to belong \textit{solely} to the source with the higher peak in
the channel autocorrelogram. If the ‘pixel’ is masked then a 0
goes into the mask indicating this. The soft mask considers the
contribution of both sources when they are present. While each
spectro-temporal ‘pixel’ is still assigned to a single source the
value in the mask is now a weight on the interval [0, 1], which
indicates the relative certainty that the point belongs to the given
source.

Given the nature of the mixtures it is possible to create
masks for each source separately and to recognise the utter-
ance of the individual speakers. The first set of experiments uses
the harmonic regions belonging to each source to perform standard
missing data recognition. In conventional missing data recog-
nition the mixed source is separated into reliable and unreliable
spectro-temporal regions. In this case the reliable regions are
those dominated by the harmonic portions of the target source.
The observed energy of the unreliable regions (those dominated
by the masking source’s harmonic portions) are used as an upper
bound for the target source energy. The contribution of inhar-
monic components is effectively removed by using the missing
data ‘bounded-marginalisation’ technique [12]. This works by
integrating over all possible values of the inharmonic energy
appropriately marginalising the probability distributions when
computing the HMM-state likelihoods.

Experiments compared results using either harmonic re-
gions generated according to section 2.1 (estimated pitch) or
harmonic regions generated using ‘\textit{a priori} pitch’. The \textit{a priori}
pitch masks were created by replacing the per frame pitch esti-
mates employed in section 2.1 with estimates obtained by track-
ning the pitch in the unmixed signals using Snack [13] (an open
source version of ESPS/waves++). With an ideal pitch tracking
algorithm this is the mask that would be generated.

3.3. Combining Harmonic and Inharmonic Fragments

In [9] the inclusion of inharmonic fragments provided some im-
provement in the recognition performance. A similar experi-
ment is performed here to show that the use of soft fragments
does not negate the importance of including the inharmonic
regions in the masks. The inharmonic regions were included
in the mask as a set of unlabelled fragments generated using
the watershed algorithm described in Section 2.2. The speech
fragment decoding process was employed to find the best tar-
get/masker labelling of these fragments [5]. The decoder would
then produce a better result than when using the harmonic re-
gions alone if it is able to correctly identify the fragments which
belong to the target source. As before, the system employed
harmonic regions derived either from estimated pitches or from
\textit{a priori} pitches.

4. Results and Discussion

The full set of experimental results for the 0 db mixed gender
test utterances are shown in Tables 1 - 2. Results are shown
separately for both the male and female utterances.

<table>
<thead>
<tr>
<th>Mask</th>
<th>Harm</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>97.1</td>
<td>86.8</td>
</tr>
<tr>
<td>F</td>
<td>95.9</td>
<td>82.4</td>
</tr>
<tr>
<td>Overall</td>
<td>96.5</td>
<td>84.6</td>
</tr>
</tbody>
</table>

Table 1: Percentage recognition accuracy for different sets of
discrete fragments: Ap - \textit{a priori} Masks, Harm - Harmonic
regions, Combined - Harmonic regions plus inharmonic frag-
ments. Results shown for harmonic regions generated from \textit{a
priori} (Ap.P) and estimated pitches (Est.P). Accuracy is shown
for male M and female F speakers plus the average of both,
(Overall)

<table>
<thead>
<tr>
<th>Mask</th>
<th>Fuzzy Harm 25</th>
<th>Fuzzy Comb 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>97.1</td>
<td>89.9</td>
</tr>
<tr>
<td>F</td>
<td>95.9</td>
<td>85.9</td>
</tr>
<tr>
<td>Overall</td>
<td>96.5</td>
<td>87.9</td>
</tr>
</tbody>
</table>

Table 2: Recognition accuracy for different sets of ‘soft’ frag-
ments: Ap - \textit{a priori} Masks, Fuzzy Harm 25 ms and Fuzzy
Comb 25 ms - Harmonic-only regions and Combined harmonic
regions and inharmonic fragments, respectively; both are gen-
erated using a 25 ms analysis window in the autocorrelation.

The \textit{a priori} results in the first column show the perform-
ance achievable with ideal binary masks. This overall perfor-
nance of 96.5% compares to a baseline system using MFCCs
(with deltas, accelerations and energy values) that achieves only
52.9% accuracy.

A comparison of the results for soft and discrete harm-
onic regions shows an improvement from an overall accuracy
of 61.4% with a discrete mask (table 1) to 69.8% for soft masks

2643


Table 3: Recognition accuracy for ‘soft’ fragments (harmonic only and combined) with a 35 ms analysis window.

<table>
<thead>
<tr>
<th></th>
<th>Fuzzy Harm 35</th>
<th>Fuzzy Comb 35</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>97.1</td>
<td>90.7</td>
</tr>
<tr>
<td>F</td>
<td>95.9</td>
<td>87.3</td>
</tr>
<tr>
<td>Overall</td>
<td>96.5</td>
<td>88.9</td>
</tr>
</tbody>
</table>

(2.6). This highlights the need for introducing a measure of certainty into the masks used for recognition.

The greater performance achieved using a priori pitch masks (87.9%) indicates how much could be gained by further improving the pitch estimation algorithm. This observation inspired further tuning of the pitch tracking algorithm. It was noted that the 25 ms autocorrelation analysis window is often too short to capture two complete pitch periods of some male speakers. Increasing the analysis window to 35 ms led to a further small improvement in performance with recognition accuracy increasing a further 1.9% from 69.8% to 71.7% (table 3).

It is of interest to note that the harmonic regions of the female masks had significantly more ‘soft’ spectro-temporal ‘pixels’ than the harmonic regions belonging to the male source’s mask. It was found (on average) that 36% of the harmonic ‘pixels’ in the female harmonic mask had a measure of uncertainty compared to 9% for the male masks. This is most likely symptomatic of the situation where a female source had a pitch roughly twice the pitch of the male source. In such a scenario a strong harmonic of the male source can boost the peak of the female source in the summary autocorrelogram which can lead to incorrect assignment of that point to the female source in the discrete mask. A hard decision will incorrectly assign the spectro-temporal point to the female source. When a soft decision is employed the assignment will be similar, but the decoder still has an opportunity to decide if the female source was indeed active at that point. This is reflected in the recognition results, which show a greater increase in accuracy with the introduction of soft fragments for the decoding of the female source.

The addition of inharmonic fragments produces improvements of 5.5% and 2.0% for the discrete masks and fuzzy masks (35 ms), respectively. A final recognition performance of 73.7% for the combined masks shows that including inharmonic regions in the masks is necessary to exploit all the information available in the signal.

The overall recognition accuracy of 73.7% compares favourably with the previous results reported for connected-digit recognition systems at 0 db (e.g. [14]). This is encouraging given that the speech maskers used in the evaluation of the current system are highly non-stationary and provide a more challenging task than the noises used in past studies. The SFD system turns the non-stationarity of the noise to its advantage. This is interesting as there is evidence that humans also find speech in the presence of non-stationary noise more intelligible than speech in stationary noise [15].

The present system relies on there being both a male and a female source in order to separate the harmonic regions. Future work will generalise the current system to remove this assumption. One potential problem is that the pitch tracks of two matched gender speakers are likely to overlap and ‘confuse’ a pitch tracker. This highlights the difficulty that the current method would face if an attempt was made to track the pitch of an individual source across the length of the entire utterance. In the mixed-gender case the pitch track of each source is widely separated in frequency and easily grouped across time; this is not possible in the matched gender case. However, as the pitch of each speaker varies smoothly, by tracking pitch estimates across time, short pitch track segments can be located. By matching these pitch segments to within channel pitch estimates, fragments of the harmonic component of each source can be isolated. These fragments, like the inharmonic fragments, would not have a target/masker label attached but would be labelled during the speech fragment decoding process.

5. References