RECOGNISING SPEECH IN THE PRESENCE OF A COMPETING SPEAKER USING A 'SPEECH FRAGMENT DECODER'

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

This paper addresses the problem of recognising speech in the presence of a competing speech source. A novel two stage approach is described. A spectral representation is first divided into a set of spectro-temporal fragments where each fragment is believed to be due to a single acoustic source. An unknown subset of these will be due to the target speech source. The standard ASR search is then extended to find the most likely combination of speech model sequence and fragment subset. The technique is tested with a fragment generation stage using pitch information to locate harmonic energy components, and image processing techniques to segment the inharmonic regions of the spectrogram. The system achieves an accuracy of 65.1% on a 0 dB simultaneous connected digit sequence task with cross-gender mixtures. Extension of the technique to handle matched-gender utterances is discussed.

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 is emerging that 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 [5, 6, 7]. The computational modelling of this process, Computational Auditory Scene Analysis (CASA), has led to systems which attempt to separate sound signals by exploiting 'primitive' properties of the acoustic signal. Features such as harmonicity and onset time are used to try and form partial descriptions of separate sources by essentially clustering elements of the time-frequency representation (time-frequency 'pixels') into larger units. It is only possible to recover partial descriptions of the individual sources as inevitably some time-frequency regions of each source will be masked by other sources. However, recent work has shown that traditional speech recognition techniques can be adapted to handle such 'missing data' [8]. Due to the large redundancy of the speech signal, these missing data speech recognition systems can achieve surprisingly good performance despite severe masking.

Although missing data techniques have met with success in a range of conditions there exists an unsolved problem with the approach. The missing data theory describes what to do when you know which spectro-temporal regions are masked, but it does not say how the location of the masked regions can be estimated. Attempts to estimate the mask using primitive CASA have only succeeded under certain conditions. Indeed, it is well understood that humans employ a mixture of primitive bottom-up processes and top-down hypothesis-driven processes when perceptually organising acoustic mixtures containing complex signals such as speech [5]. So, it would seem natural that knowledge of the speech signal (i.e. the same type of knowledge employed by ASR systems) is involved in the mask estimation itself. This is the motivation for the speech fragment decoder (SFD) model illustrated in Figure 1 [9]. Rather than 'first identify the speech regions, then use speech models to recognise the speech,' in the SFD model the identification of the speech regions and the recognition proceed in parallel and with access to a common set of speech models. In this model, the role of primitive CASA is confined to that of identifying a set of so-called 'coherent' spectro-temporal fragments (spectro-temporal regions in which the energy is due to a single acoustic source). A modified ASR-style decoder is then employed to simultaneously label these fragments as foreground/background and find the correct word sequence to model the speech.

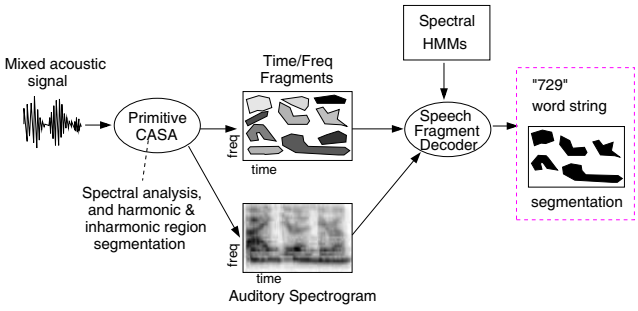

Fig. 1. An overview of the speech fragment decoding system. Bottom-up processes are employed to locate 'coherent fragments' and then a top-down search with access to speech models is used to search for the most likely combination of fragment foreground/background labelling and speech model sequence.

This work was funded by EPSRC grant GR/R47400/01

Previously the decoder has been tested in conditions in which fragment generation could be implemented using straightforward techniques. For example, it has been tested using near stationary noise backgrounds with occasional impulsive noise intrusions [9]. In the current work we study the case of speech mixed with speech at 0 dB. For a number of reasons this task is a particularly challenging - even for humans. The speech masker is highly non-stationary and so cannot be removed using simple adaptation schemes. The target and masker utterance have the same signal level, which makes them highly confusable [10]. The utterances are designed to start and end simultaneously so there is no lead or lag time in which one speaker can be heard in isolation. The signals are monaural, so common multiple microphone blind source separation techniques cannot be employed [1]. A connected digit task is employed, so high level linguistic cues are also minimised - both speakers are employing the same vocabulary and their is no grammar to constrain the possible word sequences.

The structure of the remainder of the paper is as follows. Section 2 gives a functional description of the operation of the speech fragment decoder. For the theoretical development and implementation details see [9]. Section 3 describes the novel techniques introduced in the current work to generate the fragments over which the decoder searches. Experiments conducted with simultaneous speakers uttering digits sequences are described in Section 4. In Section 5 results are presented and discussed.

2. SPEECH FRAGMENT DECODING

Consider a time-frequency representation of a speech source in the presence of one or more competing sound sources. In some spectro-temporal regions the level of the speech energy will be relatively undisturbed by the competing sources, in other regions the speech energy will be masked by a more energetic source. Missing data speech recognition systems exploit this fact to achieve robust recognition performance in noisy conditions [8]. However, as part of their input they need a binary 'mask' indicating in which regions the speech is reliable and in which it is masked. In some conditions such masks can be estimated using simple techniques (e.g. stationary background estimation). However, it is hard to see how masks can be readily generated for speech in the presence highly non-stationary noise maskers such as a competing speaker.

In contrast to missing data techniques the speech fragment decoder does not require the foreground/background segmentation to be performed a priori, rather it builds a description of this segmentation as part of the recognition process. The decoder starts with a set of coherent fragments - spectro-temporal regions that are believed to be dominated by a single sound source. There necessarily exists a unique subset of these fragments that forms a complete description of the unmasked portion of the target source. If the correct subset was known then recognition could proceed using standard missing data techniques. As the correct subset is not known, the decoder performs an algorithm that is equivalent to having separately conducted a missing data decoding using the mask generated by every possible subset of fragments, and then selecting the overall most likely decoding. Exact techniques for achieving this without a combinatorial explosion in the number of hypotheses are described in [9]. As the fragment selection is effectively conducted in parallel with the word sequence search, the foreground/background segmentation and the speech recognition mutually support each other.

The success of the speech fragment decoding technique depends heavily on the quality of the fragments being supplied as input. If the input is under-segmented and the fragments are not coherent then energy from two sources appears in a single fragment and that fragment cannot be correctly labelled as either foreground or background. Alternatively, with over-segmentation, although the fragments may be coherent some of the constraint imposed by having a small number of fragments is lost. Furthermore, over-segmentation can unnecessarily widen the search space which, depending on the recognition task, may necessitate increased pruning and hence result in decreased search accuracy.

3. FRAGMENT GENERATION

Fragment generation proceeds in two stages. The first stage identifies and groups the harmonic energy regions of each source on the basis of their fundamental frequency. This stage can proceed using techniques that have been well-established in previous source separation systems [6, 11]. However, using knowledge of harmonic energy alone renders the identification of unvoiced phonemes unreliable. So the second stage takes the remaining *inharmonic* information and fragments it into regions of high energy that appear to be well separated in time and frequency.

3.1. Harmonic Region Extraction

The harmonic region extraction stage identifies spectro-temporal regions that have a common fundamental frequency (pitch). It is based on a recent autocorrelation method for robust detection of multiple pitches within a mixture [12]. The signal, sampled at 16 KHz, is passed through a 64 channel gamma tone filter bank spaced equally on an equivalent rectangular bandwidth (ERB) rate scale with centre frequencies between 40Hz and 4000 Hz. The signal is then framed using a 25 ms window with a 10 ms frame shift. For each frame autocorrelations are computed within each frequency channel. For low frequency channels the autocorrelation is computed directly from the filter output. Since the harmonics for high frequency channels are known to be generally unresolved, the autocorrelation for these frequencies is computed from the envelope of the filter response. A summary autocorrelation is then computed by summing the autocorrelation functions for the low frequency channels. The lags at which the peaks in the summary fall above an experimentally determined threshold 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, or candidates, 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 in a single channel, then the one that has the highest amplitude is chosen.

The above process is performed on each frame to effectively group channels across frequency if they share the same pitch. A complete system would require a separate tracking stage to grow fragments across time. The current work sidelines the pitch tracking problem by focusing on mixtures of male plus female speech for which the pitches are well separated. In this condition the male harmonic region and the female harmonic region 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 is an extension of the present work which is discussed in Section 5.

3.2. Inharmonic Fragment Generation

The segmentation of the inharmonic energy exploits the fact that most of the energy in a speech signal appears in concentrated time-frequency regions. As a consequence, when multiple sources are present, energy regions of the individual sources only partially overlap, so high-energy features of the individual sources often appear as separated peaks of energy. These peaked regions can be segmented using an algorithm employing the watershed transform (commonly employed in computer vision for segmenting gray scale images [13]). Essentially the algorithm searches for intensity peaks and troughs in the image then groups regions that fall between the troughs. To avoid over-segmentation major troughs are emphasised and insignificant ones are removed. Hu and Wang [11] have recently demonstrated the potential of similar image processing techniques for segmenting auditory spectrograms.

After the watershed segmentation, inharmonic fragments that start at the same time frame can be regrouped because sound elements that share a common onset time are likely to belong to the same source. Grouping by common onset reduces the total number different fragments occupying a single time frame which greatly reduces the search space of the decoder (Figure 2).

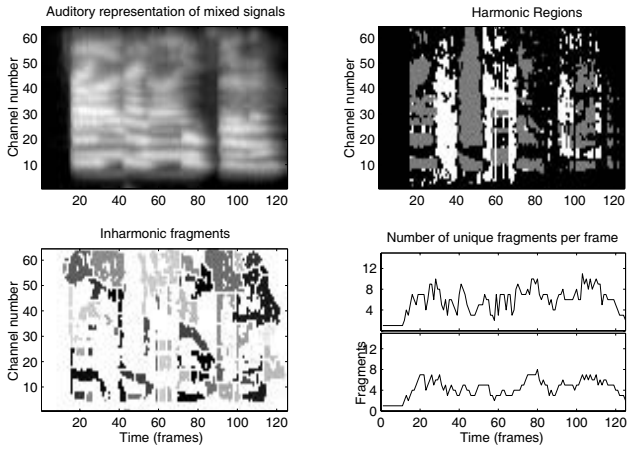

Fig. 2. Example of fragments generated from a mixture of male and female speech. Top left, mixed speech; top right, harmonic regions; bottom left, inharmonic fragments. The bottom right plot shows the number of unique fragments found at each time frame before common onset grouping (top) and after (bottom).

4. EXPERIMENTS

4.1. Test Data and Model Training

The system was tested using monaural mixtures of two simultaneous speakers (mixed gender) uttering sequences of digits mixed at 0 dB. The 1001 clean utterances from test set A of the Aurora 2 corpus were used to create the test data [14]. An end-point detection algorithm was employed to remove initial and final silences [15]. The end-pointed utterances were then ordered by length and each signal was paired with its neighbour to create 1000 pairs. Of these, the 484 pairs which had mixed-gender were used to form the test set. Mixtures were constructed by adding the signal pairs in the time domain. The shorter of each pair was padded with zeros (equally at either end) to match the size of the longer signal. The average difference in length was 0.3% with only 35 pairs having a difference of greater than 1%.

Acoustic vectors were formed by filtering with a 64 channel gammatone filter bank with centre frequencies equally spaced on an ERB scale from 50 Hz to 3850 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a 1st order filter (with an 8 ms time constant) and sampled at a frame rate of 10 ms. Cube root compression was then applied to the envelope values.

Whole word gender dependent HMMs were trained using the 4220 utterances of each gender in the Aurora clean training set. 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.

By comparing the time frequency representations of the *unmixed* signals it is possible to generate a mask of the spectro-temporal regions of the mixture in which the target is undisturbed by the masker. When missing data recognition is performed using these 'a priori' masks the results shown in the first column of Table 1 are obtained. This result represents the upper bound on what is obtainable with perfect source/target segregation.

4.2. Employing Harmonic Regions Alone

The first experiment used standard missing data techniques to examine recognition performance using harmonic information alone. The features in the spectro-temporal region dominated by the target speaker's harmonic component are treated as reliable. The features in the region dominated by the masker's harmonic component are masked. The observed masker energy in these regions is as an upper bound for the target source energy. This constraint is modelled using the missing data 'bounded-marginalisation' technique [8]. The contribution of inharmonic components is effectively removed by appropriately marginalising the probability distributions when computing the HMM-state likelihoods (i.e. integrating over all possible values of the inharmonic energy).

Experiments compared results using either harmonic regions generated according to Section 3.1 (estimated pitch) or harmonic regions generated using 'a priori pitch'. The a priori pitch masks were created by replacing the per frame pitch estimates employed in Section 3.1 with estimates obtained by tracking the pitch in the unmixed signals using Snack [16] (an open source version of ESPS/waves+). The 'a priori pitch' masks results are an estimate of the performance that could be gained by having an ideal pitch estimation algorithm.

4.3. Adding Inharmonic Fragments

The second experiment was similar to the first except the inharmonic regions, rather than being ignored, were included as a set of unlabelled fragments generated using the watershed algorithm described in Section 3.2. The speech fragment decoding process was employed to find the best target/masker labelling of these fragments. There is extra information in the inharmonic regions so the decoder should produce a better result than when using the harmonic regions alone, as long as it is able to correctly identify which fragments belong to the target source. As before, the system

employed harmonic regions derived either from estimated pitches or from 'a priori' pitches.

5. RESULTS AND DISCUSSION

The full set of experimental results are shown in Table 1. Results are shown separately for both the male and female utterances. Using the harmonic region alone produces a recognition accuracy of 59.1% and 64.5% for female and male utterances respectively. The results using a priori pitch suggest that these results could be increased by an absolute 5% to 10% by improvements to the pitch estimation algorithm. With the harmonic-only results as a baseline, it can be seen that the use of the speech fragment decoder to incorporate information from the inharmonic regions increases overall recognition accuracy from 61.8% to 65.1%.

	Ap	Harm		Harm+Inharm	
		Ap.	*Est.*	*Ap.*	*Est.*
M	97.1	69.4	64.5	79.6	66.3
F	95.9	69.3	59.1	79.3	63.8
Overall	96.5	69.4	61.8	79.5	65.1

Table 1. Recognition accuracy for different sets of fragments: *Ap* - a priori Masks, *Harm* - Harmonic regions, *Harm+Inharm* - Harmonic regions plus inharmonic fragments. Results are reported for both a priori pitch and estimated pitch harmonic regions. Results for both the male **M** and female **F** utterances are recorded. **Overall** is the average of **M** and **F**.

The modest improvement bought by the inclusion of the inharmonic regions is encouraging, but the final full-system result is still a long way short of the a priori mask score of 96.5% which indicates what could be achieved if the target and masker regions were perfectly segmented. A large part of this difference may be due to poor estimation of the harmonic regions and error introduced by the naive clustering approach. Note, using an a priori pitch estimate the full system recognition accuracy improves from 65.1% to 79.5%. However, even when using the a priori pitch estimates some channels will not necessarily be assigned to the correct source. For example when the female source has roughly double the pitch of the male source the coincidence between autocorrelation peaks caused by the two signals can make the dominant peak appear to be due to a less energetic male source.

To put the overall recognition accuracy of 65.1% into perspective it may be compared to the performance of various connected digit recognition systems that were evaluated using the Aurora corpus at a special session of Eurospeech 2001. The evaluation employed the same connected digit data artificially mixed with a variety of environmental noises – all of which were more stationary than the speech maskers employed in the current study. Our current result is closely comparable to the best of those reported for systems designed to be trained on clean speech (e.g. [17]), which is highly encouraging given the challenging nature of the speech plus speech task.

Future work will aim to generalise the current system to deal with matched gender mixtures for which the voiced regions of the two speakers can not be separated by simple clustering. 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 decoding within the SFD framework.

6. REFERENCES

[1] A.J. Bell and T.J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1004–1034, 1995.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–588, 1994.

[3] A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise.," in *Proc. ICASSP '90*, Alburquerque, NM, Apr. 1990, pp. 845–848.

[4] E.C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[5] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[6] M.P. Cooke, *Modeling Auditory Processing and Organization*, Cambridge University Press, Cambridge, U.K., 1993.

[7] A.J.W. van der Kouwe, D.L. Wang, and G.J. Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE Transactions on Speech and Audio Processing*, pp. 229–241, 2001.

[8] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[9] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, In press.

[10] D.S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers.," *J. Acoust. Soc. Am.*, vol. 109, pp. 1101–1109, 2001.

[11] G. Hu and D.L. Wang, "Auditory segregation based on event detection," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[12] M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Signal Processing*, vol. vol. 11, pp. 229–241, 2003.

[13] J.B.T.M. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol. 41, pp. 187–228, 2001.

[14] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP '00*, 2000, vol. 4, pp. 29–32.

[15] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[16] K. Sjolander, "The snack sound toolkit version 2.2b1, http://www.speech.kth.se/snack/," 2002.

[17] J.P. Barker, M.P. Cooke, and P.D. Green, "Robust ASR based on clean speech models: An evalutation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech '01*, Aalborg, Denmark, 2001.