A ‘speechiness’ measure to improve speech decoding in the presence of other sound sources

Ning Ma and Phil Green

Department of Computer Science, University of Sheffield, UK
{n.ma, p.green}@dcs.shef.ac.uk

Abstract

When speech is corrupted by other sound sources certain spectro-temporal regions will be dominated by speech energy and others by the noise. Listeners are able to exploit these cues to achieve robust speech perception in adverse conditions. Inspired by this perception process a ‘speech fragment decoding’ technique has shown promising robustness when handling multiple sound sources. This paper proposes an approach to estimating ‘speechiness’ – a degree of confidence that a spectro-temporal region is dominated by speech energy – using the modulation spectrogram. This additional knowledge is employed to steer the decoder towards selecting more reliable speech evidence in noise. Experiments show that the speechiness measure is capable of improving recognition accuracies in various noise conditions at 0 dB global signal-to-noise ratio.

Index Terms: speechiness, modulation spectrogram, speech fragment decoding, robust speech recognition

1. Introduction

In everyday listening conditions speech is naturally mixed with competing sounds, e.g. background music in a restaurant or speech babble in a bar. While listeners are adept at perceiving speech in such conditions, the noise imposes a great challenge to automatic speech recognition (ASR) systems: their performance still remains far behind that of human speech recognition (HSR). Listeners are able to segregate a mixture into perceptual packages based on source characteristics, allowing whatever package is of interest at the time to be selectively attended to. Bregman describes this process as ‘auditory scene analysis’ (ASA) [1]. This analytical ability is largely attributed to two properties of speech energy distribution in the spectro-temporal plane. First, speech energy is concentrated in local regions and these high energy regions are typically sparsely distributed [2]. This means that even at 0 dB global signal-to-noise ratio (SNR) there will be regions that are totally dominated by the noise energy, and other regions where the energy from other sources can be ignored compared to the speech energy. Secondly, speech has a redundant encoding such that it remains intelligible even if a large part of the speech spectrum is removed [3]. The redundancy essentially allows listeners to perceive speech in noise based on relatively sparse information.

The robustness of HSR has motivated extensive research into developing a new generation of robust ASR systems. For example, ‘missing-data ASR’ [4] adapts the conventional probabilistic ASR formalism to deal with partially corrupted data. The ‘missing data mask’ – a time/frequency (T/F) map of binary values – is used to label T/F pixels as being either ‘reliable’ or ‘unreliable’. One limitation of missing-data ASR as well as many other robust ASR approaches is that the processes of sound source segregation and recognition are decoupled. They adopt ‘left-to-right’ processing in which source segregation is employed as a ‘front-end’ to ASR. The prior knowledge available in recognition models is ignored when the segregation hypothesis is formed and the recognition process lacks the ability to recover segregation errors. Human speech perception, however, is believed to be interactively governed by bottom-up grouping rules that exploit common characteristics of the acoustics and top-down constraints employing the knowledge of familiar patterns [1].

Inspired by this ASA account of auditory organisation, Barker et al. [5] proposed a ‘speech fragment decoding’ (SFD) technique which combines segregation and recognition in a tightly coupled process. Primitive grouping techniques (e.g. multipitch analysis [6]) are employed to segregate the spectro-temporal plane of the mixture into a set of fragments – local spectro-temporal regions dominated by energy from individual sound sources. Statistical model-driven processes then employ speech models to simultaneously search for the most likely word sequence and foreground/background segregation that best match the recognition models. The SFD system has demonstrated promising robustness when handling multiple sound sources. For example, in a simultaneous speech recognition task Barker et al. [7] showed that SFD is capable of achieving low word error rates comparable to those of listeners. The strength of SFD is that it is designed to operate without strong assumptions about the interfering noise. It considers each fragment as being part of either the speech foreground or the noise background with equal probability. However, there are two additional factors to be considered:

• The constraints encoded in recognition models may not always be sufficient to select correct fragments [8] and extra top-down information may be required.
• Although an accurate noise model is difficult to estimate, there is often some knowledge about the noise available.

In this study we propose a ‘speechiness’ measure for a given fragment – a value in [0, 1] expressing a degree of confidence that the fragment is part of the speech foreground. The measure can be used to steer the decoder towards selecting more reliable speech evidence in adverse conditions. In the next section we describe techniques to measure and incorporate speechiness. Section 3 describes the speech and noise materials used in this study. In Section 4 a modulation filtering technique is proposed as a speechiness measure. Section 5 presents experiments and results and Section 6 concludes with future research directions.

2. Fragment Decoding with Speechiness

Given a set of fragments ASR can be performed using the SFD technique [5]. Each fragment may be variously labelled as part
of either the speech foreground or the noise background. A foreground/background segregation hypothesis is defined by a unique selection of fragment labels which can be represented by a missing-data mask. The decoder evaluates the likelihood of each hypothesised word sequence giving a segmentation hypothesis employing missing-data techniques. Barker et al. [5] demonstrated that the search process can be implemented efficiently as illustrated in Fig. 1 where three fragments (shown using the shaded regions) are being decoded. Each time a new fragment starts, all ongoing segregation hypotheses are split so that in each pair one hypothesis labels the fragment as speech while the other assigns it to the background. When a fragment ends, pairs of hypotheses are merged if their labelling only differs with regard to the fragment. At this time if the speechiness \( c_f \) of the ending fragment \( f \) is estimated, then in each pair \( c_f \) is added to the hypothesis that labels the fragment \( f \) as speech and \( 1 - c_f \) to the other. This process continues until the end of the utterance.

Figure 1: The evolution of parallel segregation hypotheses with speechiness \( c_f \) being applied. The shaded dots indicate which ongoing fragments are being treated as part of the speech foreground (adapted from [5]).

The speechiness values range in \([0, 1]\). Values 1 and 0 respectively represent that the fragment is definitely part of the speech foreground and the background. Value 0.5 gives equal weight to either hypothesis. Because of the high dimensionality of the feature vectors typically used, a scaling factor is needed to control the impact of \( c_f \). A natural candidate is the fragment size, \( s_f \), i.e. the number of T/F pixels included in the fragment \( f \), since the observation probability calculation involves the same amount of data. Therefore in log-domain \( s_f \cdot \log(c_f) \) and \( s_f \cdot \log(1 - c_f) \) are applied to each hypothesis pair.

The basic approach to the speechiness measure is to use a set of features extracted from fragments to estimate a speech model, \( M_s \), and a background model, \( M_b \). The background model can be trained using various non-speech sounds. Given the feature \( x_f \) extracted from an unknown fragment \( f \), its speechiness \( c_f \) can be derived from the posterior probability of the speech model \( M_s \) (e.g. using a sigmoid function):

\[
c_f \leftarrow \frac{P(M_s|x_f)}{P(x_f)}
\]

If we assume \( P(M_s|x_f) + P(M_b|x_f) = 1 \) and the priors \( P(M_s) = P(M_b) \), then using the Bayes’ rule Eq. 1 becomes:

\[
c_f \leftarrow \frac{P(x_f|M_s)P(M_s)}{P(x_f)} = \frac{P(x_f|M_s)}{P(x_f|M_s) + P(x_f|M_b)}
\]

In this case estimating speechiness is similar to solving a speech/non-speech classification task on which there is substantial literature. For example, Scheirer and Slaney [9] proposed using 13 features to discriminate speech from music.

If the background model is not available, speechiness can be measured by converting values derived from the fragment features to scores in the range of \([0, 1]\). These features, however, should be designed to reflect the characteristics of speech fragments so that noise fragments will score low.

### 3. Speech and Noise Material

600 utterances were randomly drawn from the Grid corpus [10]. The Grid corpus consists of utterances spoken by 34 native English speakers, which are short sentences (around 2.2 secs) of the form: <COMMAND> <COLOUR> <PREPOSITION> <LETTER> <NUMBER> <ADVERB>, e.g. ‘lay white by L please’. The utterances were normalised to have equal root-mean-square (rms) energy and sampled at 25 kHz. Six types of noises with various characteristics were selected:

1. **Violins**: Vivaldi Spring mvt 1 Allegro, harmonic source, most energy in high frequency bands
2. **Piano**: Chopin Nocturne op 9 no 2, harmonic, most energy in low frequency bands
3. **Singing voice**: female vocal solo with piano accompaniment, harmonic, mostly overlapping speech energy
4. **Drums**: from [8], fast rhythms, mostly overlapping speech energy
5. **Speech babble**: from NOISEX-92 [11], unstationary, mostly inharmonic and overlapping speech energy
6. **Factory noise**: from NOISEX-92, a stationary background with highly unpredictable components such as hammer blows etc, inharmonic full band noise

All noise signals were resampled to 25 kHz and normalised to have target rms of 0.05. Noisy mixtures were produced by adding each of the noise signals to the 600 test utterances at 0 dB global SNR. The starting point in the noise signals was randomly set each time. We are particularly interested in the 0 dB SNR because in this condition the level difference cue that can be exploited by SFD to recruit correct fragments is minimised.

Cochlear frequency analysis was simulated via a bank of 64 overlapping gammatone filters with centre frequencies spaced uniformly between 50 Hz and 8 kHz on the ERB scale. The instantaneous Hilbert envelope at the output of each filter was smoothed using a first order filter with an 8 ms time constant, sampled at 10 ms intervals, and finally log-compressed to produce an auditory spectrogram (see Fig. 2a). For each mixture a set of oracle fragments was generated by making use of priori knowledge of the pre-mixed signals. Using such oracle fragments allows us to study the fragment selecting problem in isolation of the fragment generating problem. An example of fragments generated in this way for a mixture of speech and speech babble is shown in Fig. 2b.

### 4. Speechiness by Modulation Filtering

Speech has characteristic low-frequency amplitude modulations associated with the syllabic rate (around 4 Hz). Modulations above 16 Hz are not essential for human speech perception [12]. Kingsbury et al. [13] therefore proposed a feature representation which emphasises this temporal structure – the ‘modulation spectrogram’ – for ASR in adverse conditions. Palomäki et al. [14] used a similar technique to identify regions dominated by direct speech rather than reverberation. In this study the modulation filtering technique is employed as a speechiness
measure. In brief, a modulation filter is applied to each frequency band and periods where significant energy gets through are identified. The energy levels are then averaged over pixels in a fragment to judge its speechiness.

Analysis on Grid utterances shows that the word rate has a peak around 4 Hz \(^1\) (shown in Fig. 3a). The statistics were calculated from word duration samples obtained using forced-alignment. We therefore design a bandpass modulation filter, \( h(t) \), with a pass-band of 2.5–6.67 Hz to emphasise the 4 Hz energy. Fig. 3b shows its frequency response. The spectral energy in each frequency band is filtered with \( h(t) \) and the peak level over all bands is set to 0 dB. Levels more than 40 dB below the peak are set to −40 dB. The process produces a representation (shown in Fig. 2d) similar to the modulation spectrogram [13].

\[ c_f = \frac{1}{1 + \exp(-\alpha(e_f - \beta))} \]  

where \( \alpha = 0.02 \) is the sigmoid slope and \( \beta = -27 \) is the sigmoid centre. The parameters were derived via the equal error rate (EER) analysis in a classification experiment using a separate development dataset and fixed for all the noise conditions.

5. Experiments and Discussions

5.1. SFD system setup

The speech recognition task is to identify the letters and digits in each utterance. Speaker-independent word-level HMMs were trained using 500 clean utterances per speaker (different from the test set). Each word was modelled using two states per phoneme with 16 diagonal-covariance Gaussian mixture components per state and a left-to-right no-skip topology. The SFD system employed spectral features (as shown in Fig. 2a) which are supplemented with their temporal derivatives to form 128-dimensional feature vectors.

5.2. Results

Oracle fragments were used by the SFD system in 5 experiments: a) ‘SFD baseline’ – SFD with no speechiness; b) ‘correct segment’ – SFD forced to recruit all true speech fragments and no noise fragments; c) ‘speechiness 1.0/5’ – forced to recruit all true speech fragments (given speechiness 1) but treat noise fragments with equal probability (given speechiness 0.5); d) ‘speechiness 0.5/0’ – forced to miss out true noise fragments (given speechiness 0) but treat speech fragments with equal probability (given speechiness 0.5); e) ‘modulation filter’ – fragments given speechiness estimated using the modulation filtering technique. Experiments (a–b) examine the performance difference due to fragment selecting errors and (b) will give the upper bound of SFD performance. Control experiments (c–d) investigate the decoder’s ability to select fragments. The fragment identities were revealed using prior knowledge so that controlled speechiness can be assigned. In experiment (e) no prior knowledge of fragment identities were used.

Fig. 5 shows overall recognition accuracies of the two keywords in various noise conditions (SNR = 0 dB). The baseline SFD system produced significantly lower accuracies than those using the correct segmentation. A similar observation was reported in [8]. Control experiments (c–d) suggest that this performance drop is largely due to failure to recruit enough speech fragments rather than missing out noise fragments. When

---

\(^1\) This is approximately the syllabic rate since most words in the Grid corpus are single-syllable.
forced to employ all the true speech fragments, the decoder showed little difficulty discarding noise fragments, which leads to accuracies close to the human measure in most noise conditions. It did not select enough speech evidence even when all the noise fragments were forced to be assigned to the background.

When employing speechiness estimated using the modulation filtering technique, SFD produced results close to the top performance. The improvement is particularly large in noise conditions (such as ‘piano’) which have a temporal structure significantly different from that of speech. In these conditions the speechiness measure is more reliable. Note the speechiness measure does not have to be 100% accurate as it is presented as a degree of confidence. Inappropriate values can still be offset if the acoustic evidence strongly matches recognition models.

6. Concluding Remarks

SFD [5] provides a general framework to couple source segregation and recognition. In this study we integrate a speechiness measure into the SFD framework to bias the decoder towards selecting fragments that are more likely to be part of the speech source. A modulation filtering technique which emphasises the characteristic low-frequency modulation energy of speech appears to be a good speechiness measure. Recognition experiments show that the speechiness measure can help the decoder employ more reliable speech evidence and therefore produce significant accuracy improvement.

In a broader view the speechiness can be seen as a general weight given to each fragment representing a confidence of being part of the target foreground. The proposed technique provides an approach to exploiting extra top-down constraints for sound source segregation and recognition. In [5] SFD bases the selection of fragments only on recognition models, which models speech spectra. However, there is plenty of other information about the fragments that is not available in these models, e.g., two fragments from different directions may both match speech models well but cannot both belong to the target source. Long-term properties, such as the modulation spectrogram employed in this study, is also an example. Solutions to auditory scene analysis often require various constraints. In order to produce a robust estimate of such a confidence, various features, e.g., pitch and location cues, should be combined. There is also much evidence that listeners are able to quickly form simple noise models [1]. Employing noise models (not necessarily as detailed as the speech HMMs) will bring more reliable estimates, especially when present sources have many similarities.

Future work also includes investigating methods of applying speechiness measures to each segregation hypothesis which may include multiple fragments.

7. Acknowledgements

The authors would like to thank Dr. Jon Barker for valuable discussions on the SFD implementation and providing some noise signals, and Dr. Kalle Palomäki for helpful comments on the modulation filter design. This work was partly supported by the EU 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, IST-033812).

8. References