DETECTION, SEPARATION AND RECOGNITION OF SPEECH FROM CONTINUOUS SIGNALS USING SPECTRAL FACTORISATION

Antti Hurmalainen[⋆] *Jort F. Gemmeke*[†] *Tuomas Virtanen*[⋆]

[⋆] Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland
[†] KU Leuven, Department ESAT-PSI, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

ABSTRACT

In real world speech processing, the signals are often continuous and consist of momentary segments of speech over non-stationary background noise. It has been demonstrated that spectral factorisation using multi-frame atoms can be successfully employed to separate and recognise speech in adverse conditions. While in previous work full knowledge of utterance endpointing and speaker identity was used for noise modelling and speech recognition, this study proposes spectral factorisation and sparse classification techniques to detect, identify, separate and recognise speech from a continuous noisy input. Speech models are trained beforehand, but noise models are acquired adaptively from the input by using voice activity detection without prior knowledge of noise-only locations. The results are evaluated on the CHiME corpus, containing utterances from 34 speakers over highly non-stationary multi-source noise.

***Index Terms*—** Spectral factorization, speech recognition, speaker recognition, voice activity detection, speech separation

1. INTRODUCTION

Applying automatic speech recognition (ASR) in noisy environments introduces several new challenges not present in clean conditions. A fundamental problem is corruption of speech features by additive noise, which may not match to noise observed during model training. In previous work, high separation quality has been achieved by applying spectral factorisation that decomposes a noisy input spectrogram into activations of multi-frame speech and noise *atoms*, which can be acquired from training material or from the local context [1, 2, 3, 4]. We have shown that a method known as *sparse classification* (SC), which determines the phonetic content directly from the weights of activated speech atoms, can produce speech recognition results comparable to source separation followed by conventional back-end recognition [1, 5].

In previous experiments with noise atoms sampled from the neighbourhood of noisy utterances, we have used annotated speech endpointing to sample from segments known to consist of only noise. In real world applications, such information cannot be assumed to be available, thus speech activity must be estimated. In other speech recognition methods, voice activity detection (VAD) has been employed to detect speech and noise segments and to update the noise model [6].

In this work, we propose the use of SC-based methods for detecting the target utterances from mixtures containing high noise levels and occasionally overlapping non-target speech. The same framework is used for noise model updating and subsequent source separation. Speech models are acquired beforehand from training material, whereas noise models are adapted from the context.

Another topic of interest is the use of speaker-dependent speech recognition to obtain better results in both clean and noisy environments. However, the true speaker identity may not be known during recognition. We propose SC for determining the speaker identity from continuous noisy mixtures, whereafter source separation and speech recognition is carried out with speaker-dependent speech models.

The work is organised as follows: Section 2 introduces the main concepts of factorisation-based speech separation and recognition. In Section 3 we present the framework for processing continuous audio, detecting speech locations, and updating the noise model. In Section 4 we apply the algorithms to CHiME data, consisting of utterances from 34 speakers over continuous, highly non-stationary background noise. Finally, in Section 5 we draw the conclusions.

2. FACTORISATION-BASED SPEECH SEPARATION AND RECOGNITION

The methods presented here are based on representing an observed sound mixture as a linear sum of speech and noise *atoms*, each belonging to a single speaker or to background noise. The features consist of Mel scale spectral magnitudes, computed in 25 ms *frames* with a 10 ms shift. The atoms are $B \times T$ spectrogram segments, where B is the number of Mel bands and T is the number of consecutive frames in an atom. Speech and noise atoms form a *dictionary* (or *basis*). By assuming that magnitudes of multiple sources are approximately additive in the Mel-spectral domain, factorisation becomes a problem of finding non-negative *activation weights* x_l for each atom index $l \in [1, L]$ in the system, together denoted as an *activation vector* $\mathbf{x}$.

The research of Jort F. Gemmeke was supported by IWT-SBO project ALADIN contract 100049.

2.1. Convolutive spectral factorisation

An observation spectrogram $\mathbf{Y}$ ($B \times F$), where the number of frames F is larger than atom duration T, is factorised using convolutive temporal modelling and joint spectrogram estimation with overlapping segments [7]. We find the $L \times W$ *activation matrix* $\mathbf{X}$, consisting of an activation vector for all W *window indices* in the observation. We only consider windows fitting completely within the observation spectrogram. Thereby the activations of the final window index W takes place at time $F - T + 1$. $\mathbf{X}$ is obtained by optimising the estimated observation spectrogram $\mathbf{\Psi}$, modelled convolutively as

$$\mathbf{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \qquad (1)$$

Each $\mathbf{A}_t$ ($t \in [1, T]$) is a $B \times L$ matrix, containing frame t of every $B \times T$ atom in the dictionary. Operator → shifts the columns of $\mathbf{X}$ right within a $L \times F$ zero-padded matrix by $t - 1$ columns. The cost function to be minimised consists of Kullback-Leibler divergence between $\mathbf{Y}$ and $\mathbf{\Psi}$, and the sum of $\mathbf{X}$ entries weighted element-wise by a sparsity penalty matrix. The exact cost functions and iterative update rules used in our convolutive factorisation are described in [2, 5].

2.2. Source separation and sparse classification

The activation matrix can be used for source separation. Two spectrogram estimates are derived from Equation 1; a noisy speech reconstruction $\mathbf{\Psi}$ obtained by using both speech and noise atoms, and a clean speech estimate $\mathbf{\Psi}_s$ obtained by only using speech atoms and activations. The element-wise speech-to-total ratio $\mathbf{\Psi}_s / \mathbf{\Psi}$ is converted back to discrete Fourier frequency scale by multiplication from the left by a pseudoinverse of the Mel filterbank matrix, and acts as a time-varying filter for the original mixture spectrogram. It is then used to estimate speech-only features and further to synthesise separated time-domain signals [5].

To determine the speaker identity and phonetic content from speech atom activations, each speaker-dependent speech atom is associated with a $Q \times T$ *label matrix* $\mathbf{B}$. It represents the presence of each phonetic state $q \in [1, Q]$ over the atom's frame indices [1, 5]. These atom-state labels are used to calculate a $Q \times F$ *likelihood matrix*, representing phonetic state likelihoods over the whole duration of the observation. The likelihood matrix is calculated by applying Equation 1, with state label matrices $\mathbf{B}$ taking the place of the atom spectrograms. The method is known as *sparse classification*. In previous work we have used it for speech decoding [1, 2, 5]. Here the state likelihood information is used for voice activity detection and speaker identification.

3. PROPOSED SYSTEM FOR PROCESSING CONTINUOUS AUDIO

In the proposed system, continuous input audio is processed gradually using convolutive spectral factorisation, a fixed multi-speaker speech basis obtained in the training stage, and an adaptively updated noise basis. As the factorisation advances within the signal, speech activation weights and state mapping matrices are used to construct estimates of the presence of phonetic states for each speaker individually. The speaker-dependent state information is used for two purposes, speech locating and speaker identification.

3.1. Voice activity detection

We perform initial factorisation in 750-frame (7.5 s) spectrogram *blocks*. An extended Hann window function, consisting of 250 frames of fade-in, 250 frames of flat top and 250 frames of fade-out is applied to each block spectrogram. 2/3 overlap is present between blocks, so that each frame of the input is included in exactly one flat middle section. Blocks are factorised consecutively using the convolutive model described in Section 2.1 with a multi-speaker speech basis (Section 4.2) and an adaptive noise basis (Section 3.2).

Speech activations are converted into phonetic state likelihoods by using mapping matrices $\mathbf{B}$ and overlap-added over blocks. Using the initial state likelihood estimates and word-dependent *VAD weight functions* over time, a total VAD level estimate is derived for each input frame. Each word is assigned a specific weight profile over time, spanning up to 30 frames to both directions from the original frame location for temporal smoothing and utterance modelling. Based on the task grammar, the shape of weight functions depends on the role of each word in a sentence: the functions corresponding to the first and last word classes in a sentence are given negative weight before and after them, respectively. This emphasises the contrast in VAD level between target speech and its surroundings, helping to isolate test utterances from noise and non-test speech. An example of weight functions that were used in the simulations is shown in Figure 1. Word activity sums are convolved with their respective weight functions, and then summed together for the total VAD weight.

Speech-noise classification is performed using the total VAD weight over frames and on/off threshold values determined from development data. In addition, constraints can be set on the utterance duration to select candidates matching to the expected temporal profile of utterances.

3.2. Noise basis acquisition

Areas flagged as noise are sampled directly into noise atoms with a $T/2$ overlap between consecutive atoms. A threshold value is used on the spectrogram magnitude sum of segments to only store atoms with significant noise events. A noise dictionary is maintained, starting empty and acquiring new content up to a defined maximum capacity. Each noise dictionary atom is given a *significance weight*, increasing according to its activation weight in factorisation and decaying exponentially over time. Whenever newly introduced noise atoms would exceed the dictionary size, the least significant existing atoms are discarded. The latest dictionary is always used for factorisation.

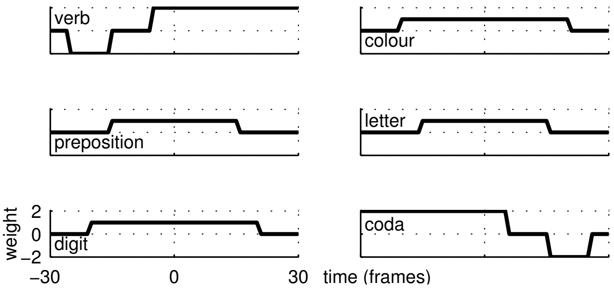

Fig. 1. VAD weight functions for each CHiME word class.

3.3. Speaker recognition

As we use a multi-speaker basis with knowledge of the speaker identity of each speech atom, a likelihood matrix can be generated for each speaker individually. For a span of frames marked as speech, we find the maximum sum of speaker-dependent state content to identify the most likely speaker. The identification result, in turn, is used for another, local factorisation pass so that only the chosen speaker's speech basis is included. By narrowing down the speech basis, the system becomes more sensitive to the chosen identity and may be able to pick the correct phonetic content even from mixtures containing other speakers. In separation-based speech recognition, the identity estimate is also used for selecting the speaker-dependent GMM model in the back-end.

4. EXPERIMENTS

4.1. CHiME data

The experiments were conducted on CHiME data, consisting of GRID command utterances mixed over highly non-stationary family household noises with simulated room reverberation response matching the noise [8]. The target utterances are from 34 different speakers and follow a linear six-class *verb-colour-preposition-letter-digit-coda* grammar ("set white in H 7 please"). A default language model is provided for recognition, employing 250 sub-word states for the 51-word vocabulary. For each speaker, there are 500 training utterances with reverberation but no additive noise. Development and test sets consist of a total of 600 utterances from all speakers together, repeated at multiple SNR levels. The noises contain a large variety of everyday sound events including appliances, impacts, music and also spontaneous speech from non-target speakers.

For this work, we use the continuous, 'embedded' CHiME sequences. In the test set, there are 16 *sessions* ranging from 27 to 87 minutes. The 600 test utterances are spread over the sessions at SNRs ranging from +18 to -6 dB at 3 dB intervals. SNRs from +9 to -6 dB belong to the official scoring set. The locations of speech in sessions are chosen in such a way that the target SNR is achieved by direct mixing without scaling. Therefore it is common for one loud segment of background noise to contain several low-SNR test utterances in succession. Conversely, there are also long noise-only sequences between the test utterances.

All 16 kHz CHiME audio was converted into $B = 40$ band Mel-scale magnitude spectrograms with 25 ms frame length and 10 ms frame shift, and equalised using a frequency band weighting curve derived from speech training material. For spectral processing, the magnitudes of left and right channels were averaged to form monaural spectrogram features.

4.2. Bases and labelling

A speech basis was created for every speaker by employing forced alignment data acquired from the CHiME HTK models. Based on the 250 sub-word phonetic states, each state in turn was modelled by placing its corresponding word instances from 300 training utterances in a $B \times T$ spectrogram window with the target state in the middle [5]. A median was taken over the instances within each time-frequency point, creating a characteristic template of the state spectrum and its typical neighbourhood. Atom length T was set to 25 (265 ms), which is enough to capture short words in their entirety, and partial content of longer words, together modelling slight variations in the pace of pronunciation. All in all, the 250 atoms of 34 speakers formed a 8500-atom speech basis. The remaining 200 training utterances from all speakers were combined and factorised using the full speech basis to learn the activation-state mapping matrices $\mathbf{B}$ with ordinary least squares regression as described in [2, 5].

An adaptive noise basis was maintained as described in Section 3.2. We used a maximum capacity of 500 atoms for sampled noise. In addition, 15 atoms were initialised randomly and updated during iteration to model unseen noise events, e.g. when the adaptive basis was empty [5]. The maximum number of atoms used in block factorisation was 9015 (8500 speech, 500 sampled noise, 15 on-line updated noise).

4.3. VAD accuracy

The VAD algorithm described in Section 3.1 was used to find utterances from CHiME sessions. A VAD weight function was given for each word class in CHiME grammar to reflect the expected speech activity profile in its neighbourhood. The functions are shown in Figure 1. On/off thresholds for total VAD level were acquired from development data and set to favour false positives over missed true utterances. To reflect the duration of CHiME utterances, a minimum length requirement of 80 frames was set for speech segments, and after 180 frames from the start of a segment it was ended as soon as the silence threshold was reached. Between these limits, temporary gaps of up to 60 frames were allowed to model short pauses in speech. Because the CHiME ground truth annotations occasionally contain excess silence, an utterance was ruled as being found in a segment for scoring if at least 40% of its duration was flagged as speech by VAD.

Speech detection results are listed in Table 1. Of the 5400 test utterances (600 for each SNR level), 5331 (98.7%) were detected successfully. 5090 were also assigned correctly to single segments, whereas 241 appeared in segments where

Table 1. Voice activity detection results: 600 utterances at 9 SNR levels, all in all 5400 utterances, exist within the continuous test sessions. 5939 speech segments were detected.

Found speech segments	True utterances
726 false positives 5090 containing 1 utterance 120 containing 2 utterances 3 containing 3 utterances	65 false negatives (misses) 5331 found in 1 segment 4 split between 2 segments
Total: 5939	Total: 5400

two or more consecutive utterances got merged. In a few cases, an utterance was split between two found segments. 726 false positives — segments with no target utterances — were also found. These mostly consisted of other speech found in CHiME background noise.

In a completely realistic scenario, the detected speech segments should be identified and recognised by themselves. In these experiments we used found segments for VAD quality evaluation and noise modelling, but annotated endpointing for speaker identification and speech recognition. The reason for this choice is that the default CHiME language model assumes tightly cropped, single utterances as its input. Passing VAD-based segments with possible silence and merged utterances would result in unpredictable back-end behaviour and problems in comparing the scores with prior ASR methods.

4.4. Speaker identification

The results for speaker identification are listed by SNR in Table 2. The identification rates of 12–18 dB utterances were between 99–100%. We notice that above 0 dB, misclassifications are rare. From 0 dB downwards, the utterances may — and often do — contain equally loud speech from non-target speakers, which may cause the maximum activity classifier to select an identity matching to the non-target speech instead of the true speaker. Misclassification of target speech to another similar sounding speaker may also take place due to corruption of spectral features. For enhancement and sparse classification, the latter kind of errors are still tolerable, whereas the former are often unrecoverable.

4.5. Speech separation and recognition

Enhanced utterances were cropped from the full session signals separated during multi-speaker block processing. Real utterance locations were also re-factorised using a single-speaker basis of both true and estimated speaker identity in turn. The latest sampled noise basis and $\lceil F/T \rceil$ on-line updated noise atoms were used in the second, local factorisation pass. Enhanced test signals, generated as described in Section 2.2, were stored for GMM-based speech recognition and measurement of signal-to-distortion ratio (SDR) of enhanced utterances in comparison to clean test files.

Table 2. Speaker identification scores (%) on the CHiME test set over SNRs. SC-based maximum state sum is used to determine the most likely identity among 34 speakers.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Correct	99.2	98.7	97.2	91.3	85.0	74.5	91.0

For enhancement-based recognition, we used the CHiME language model and multi-condition (MC) trained speaker-dependent GMMs as in [3, 5]. Models were not retrained for enhanced signals. SDR was calculated as

$$SDR_{dB} = 10 \log_{10} \frac{\sum_n s(n)^2}{\sum_n (\hat{s}(n) - s(n))^2}, \qquad (2)$$

where $s(n)$ is the clean reference signal and $\hat{s}(n)$ is the noisy or enhanced signal over sample index n [9]. Because CHiME annotations do not match perfectly to the isolated files, signals were aligned with maximum cross-correlation before measurement. Both in recognition and SDR measurement, left and right channels were averaged to form monaural signals.

Results for speech recognition are shown in Table 3. The first half displays baseline scores for the clean-trained CHiME reference models, the MC-trained models without any enhancement, and our previous results using a 250-atom sampled noise dictionary exploiting full knowledge of noise-only segments and the same speech bases as in this work [5]. In the second half, recognition results are shown for the new VAD-based noise modelling. Four different combinations are used for the choice of speech dictionaries in factorisation and for speaker-dependent GMM models used in the back-end.

The scores generally decrease as endpointing and identity information is lost, but even in the worst case where estimated identity is used for all parts, the new results surpass unenhanced, known-identity recognition by a wide margin. Interestingly, enhancement using all speakers' bases is on average better than only using the true identity. One possible explanation is that using all bases simultaneously allows wider phonetic variation, even though not all atoms belong to the target speaker. The degradation from losing identity information in separation and GMM selection reflects the misclassification rates over SNRs seen in Table 2. The largest decrements take place in the noisy end, but overall only 2.6% (absolute) loss is observed in average accuracy when true identity is wholly replaced by an estimate.

Results for SDR measurement are shown in Table 4. The first rows show SDRs for unenhanced utterances and enhancement with the earlier 250-atom informed noise modelling. Note that the nominal CHiME SNRs do not match the measured, unenhanced SDRs due to different weighting. In the second part, results for the new, self-adapting noise model are shown. Curiously, our new model produces superior separation quality, which does not translate to better ASR rates. We can speculate that the proposed noise model with long memory and adaptive atoms is able to remove more major noise events than the strictly local, informed model. Meanwhile, it may also remove crucial speech information, thus reducing

Table 3. Enhancement-based speech recognition scores (%) on the CHiME test set over different SNRs. First part shows unenhanced baseline scores for standard CHiME models and multi-condition (MC) trained models, and the latter with informed 250-atom noise modelling. The second part uses new, self-adapting noise models. Row labels denote the speech bases used for enhancement (all/true/estimated), and the speaker model used for GMM evaluation (true/estimated).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Baseline scores and informed noise modelling							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC, none	91.3	86.8	81.7	72.8	61.1	54.5	74.7
MC, inform.	93.0	91.2	90.0	85.2	79.0	72.9	85.2
Self-adapting noise, enhancement + MC recognition							
All/true	92.8	89.8	87.8	84.4	75.5	73.9	84.1
True/true	91.6	88.8	88.2	83.9	76.9	68.9	83.0
Est./true	91.4	88.8	87.8	82.6	73.8	64.4	81.5
Est./est.	91.4	88.6	87.1	81.0	72.3	62.2	80.4

the final ASR rate. Another noteworthy observation is that the compact single-speaker speech models introduce more distortions in the clean end than using all speakers' bases, but in the noisy end they manage to separate target speech better.

5. CONCLUSIONS

Spectral factorisation based methods were presented for solving three problems; voice activity detection, speaker identification, and speech separation/recognition from a continuous input. Results were evaluated using CHiME data, containing 34 speakers and household noise at SNRs from 9 to -6 dB.

98.7% of target utterances were found by estimating voice activity from speech atom activations and state labels. False positives generally consisted of non-target speech present in CHiME noise. Non-speech segments were used to update the noise model in continuous factorisation, thereby making the model completely independent of noise training data.

Activation weights of a multi-speaker basis were used to determine speaker identity among the 34 candidates. An average identification rate of 91.0% was achieved over all SNRs. Thereafter utterances were separated for GMM-based speech recognition. The new, self-adapting noise model yielded higher signal-to-distortion ratios than earlier, informed noise modelling. However, speech recognition rates decreased slightly when speaker identity was estimated. Approximately 80% average scores were still achieved after bypassing all information on speaker identity and noise locations.

The results as a whole demonstrate, how spectrogram factorisation and sparse classification can be used for several subtasks in noise-robust speech separation and recognition. We eventually hope to extend the presented work into a complete large vocabulary continuous speech recognition framework based on SC techniques.

Table 4. Measured signal-to-distortion ratios (dB) for unenhanced and enhanced CHiME test utterances over nominal mixing SNRs.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Unenhanced signals and informed noise modelling							
Unenhanced	3.7	2.5	0.3	-1.9	-4.8	-7.0	-1.2
Informed	4.4	4.1	3.8	3.5	3.1	2.7	3.6
Self-adapting noise, all/true/estimated identity							
All	8.6	7.8	6.8	5.9	4.7	3.9	6.3
True	6.9	6.4	6.0	5.5	4.9	4.4	5.7
Estimated	6.9	6.4	6.0	5.4	4.6	4.0	5.6

6. REFERENCES

[1] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[2] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, "Exemplar-based Recognition of Speech in Highly Variable Noise," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 1–5.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[4] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans, "Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 74–79.

[5] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *submitted work*, 2012.

[6] K. Demuynck, X. Zhang, D. Van Compernolle, and H. Van hamme, "Feature versus Model Based Noise Robustness," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 721–724.

[7] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[8] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Proc. ICA*, 2007, pp. 552–559.