INTEGRATING MULTIPLE OBSERVATIONS FOR MODEL-BASED SINGLE-MICROPHONE SPEECH SEPARATION WITH CONDITIONAL RANDOM FIELDS

Yu Ting Yeung, Tan Lee

Department of Electronic Engineering
The Chinese University of Hong Kong
{ytyeung,tanlee}@ee.cuhk.edu.hk

Cheung-Chi Leung

Institute for Infocomm Research
A*STAR, Singapore
ccleung@i2r.a-star.edu.sg

ABSTRACT

A single-microphone speech separation framework based on conditional random fields (CRFs) is proposed in this paper. Unlike factorial HMM, CRF does not have the conditional independence assumption on observations, thus different types of observations from the speech mixture can be integrated into the models through feature functions. Similar to factorial HMM, there is the statistical independence assumption on sources. Under this assumption, the two-source single-microphone speech separation problem can be expressed by two independent linear-chain CRFs. The separation problem becomes two pattern recognition problems, with respect to CRF models of the two sources. Experimental results show that by integrating initial separation outputs from factorial HMM with log power spectrum, fundamental frequency and speaker likelihoods of the mixture, CRF separation framework consistently improves the results from factorial HMM in terms of SNR, segmental SNR and PESQ.

Index Terms— speech separation, conditional random fields

1. INTRODUCTION

Single-microphone speech separation problem is to reconstruct the sources from only one speech mixture. In the two-source case, speech mixture y is modeled as the addition of two independent sources s_1 and s_2, i.e. $y = s_1 + s_2$. The problem is the extreme case of under-determined source separation, which is no unique solution for source reconstruction.

Model-based method is one of the approaches to the problem [1]. The sources are modeled by statistical models. Factorial HMM has been proposed for efficient computation, with the assumptions on statistical independence between the sources and conditional independence on observations [2]. Various observations from mixture and sources are useful for separation. However, they may differ in dimension, dynamic range and representation format. They may be highly dependent, for example they are originated from the same frequency spectrum. These make the integration into factorial HMM difficult.

Conditional Random Field (CRF) is a discriminative probabilistic framework [3]. It is a Markov Random Field (MRF) conditioned on the observations. It does not require the conditional independence assumption on observations. Highly correlated observations can be integrated into CRFs straightforwardly, such that the model is more descriptive than conventional HMM. Direct modeling like CRFs also express the problem more naturally. The problem statement of a two-source case is "given the observations of a speech mixture, to find the most probable sources", in contrast to "given all the source pairs, to find the most probable pair which generates the observations of the speech mixture" in a generative approach like factorial HMM.

A CRF based separation framework is proposed in this paper. The initial separation outputs from factorial HMM are integrated with observations such as log power spectrum, fundamental frequency and speaker likelihoods of the speech mixture to obtain improved separation results. The paper is organized as follows. Section 2 provides the background of speech separation and factorial HMM. The formulation of the separation problem into CRFs is discussed in Section 3. Observations extracted for CRF separation framework are described in Section 4. Experimental results are shown in Section 5. Finally, Section 6 concludes the paper.

This research is partially supported by the Project Grant from the Shun Hing Institute of Advanced Engineering, the Chinese University of Hong Kong and the General Research Fund (Ref: CUHK 414010) from the Hong Kong Research Grants Council.

2. SINGLE-MICROPHONE SEPARATION WITH FACTORIAL HMM

2.1. Signal interaction model

Let $\hat{y}_t$, $\hat{s}_{1,t}$ and $\hat{s}_{2,t}$ be the D-dimensional log power spectra of the mixture and the two sources respectively at frame t. For each frequency component d, $\hat{y}_t$ is expressed as

$$\hat{y}_t^d = \log(e^{\hat{s}_{1,t}^d} + e^{\hat{s}_{2,t}^d} + 2e^{\frac{1}{2}(\hat{s}_{1,t}^d + \hat{s}_{2,t}^d)}\cos(\theta_t^d)) \quad (1)$$

with θ_t^d as the phase difference between the two sources. Assuming θ_t^d is uniformly distributed, $\mathcal{E}(\cos(\theta_t^d)) = 0$. Hence,

$$\log(\mathcal{E}(e^{\hat{y}_t^d})) = \log(e^{\hat{s}_{1,t}^d} + e^{\hat{s}_{2,t}^d}) \approx \max(\hat{s}_{1,t}^d, \hat{s}_{2,t}^d) \quad (2)$$

by soft-maximum approximation. This is equivalent to MIXMAX model [4], which is the non-linear MMSE estimator of $\hat{y}_t$ [5].

2.2. Factorial HMM for source separation

For the two-source case, the source signals $\hat{\mathbf{s}}_\mathbf{k} = \{\hat{s}_{k,t} : t \in T\}$, $k \in \{1, 2\}$ of length T are represented by the state sequences $\mathbf{m_k} = \{m_{k,t} : t \in T\}$ of two independent Markov processes $\mathbf{M_k}$. Let $\mathbf{O} = \{O_t : t \in T\}$ be observations, the maximum likelihood estimation of the state sequences of the sources is obtained as

$$(\mathbf{m}_1^*, \mathbf{m}_2^*) = \arg\max_{\mathbf{m_1}, \mathbf{m_2}} p(\mathbf{O}|\mathbf{m}_1, \mathbf{m}_2) \quad (3)$$

by a generative approach. The decoding can be done by the Viterbi algorithm. An illustration of factorial HMM is shown as in Figure 1. At frame t, mixture $\hat{y}_t$ is emitted as the observation O_t from the

interaction of unobserved sources $\hat{s}_{k,t}$. $\hat{s}_{k,t}$ are generated from the given states $m_{k,t}$ with emission probability densities $p(\hat{s}_{k,t}|m_{k,t})$. If each density is a Gaussian distribution $\mathcal{N}(\mu_{k,t}, \Sigma_{k,t})$ with a diagonal covariance matrix $\Sigma_{k,t}$, the probability density of $\hat{y}_t$ can be approximated as $\mathcal{N}(\max(\mu_{1,t}, \mu_{2,t}), \Sigma_t)$, where $\max(\cdot)$ is an element-wise operator, $\Sigma_t^d \approx \Sigma_{1,t}^d$ if $\mu_{1,t}^d > \mu_{2,t}^d$ for dimension d, and $\Sigma_t^d \approx \Sigma_{2,t}^d$ otherwise for computational tractability. The transition probability from state $m_{k,t-1} = i$ to $m_{k,t} = j$ is U_{ij}^k.

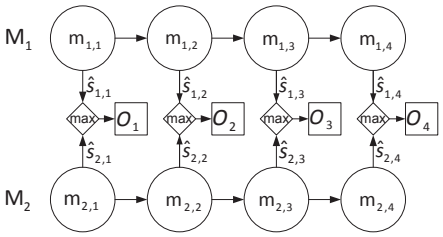

Fig. 1. Factorial HMM with MIXMAX as interaction function.

3. CONDITIONAL RANDOM FIELDS

3.1. Formulation of conditional random fields

Conditional random fields (CRFs) directly model $p(\mathbf{m}|\mathbf{O})$, the posterior probability distribution of state sequence $\mathbf{m} = \{m_t : t \in T\}$, given observations $\mathbf{O}$. Figure 2 illustrates a linear chain CRF. The vertices that are dependent of each other are connected by edges. For linear-chain CRFs, $p(\mathbf{m}|\mathbf{O})$ depends on only three types of cliques (complete sub-graphs) under Hammersley-Clifford theorem [6]: current observations and state, previous and current states, current and next states. By introducing feature functions $f_i(\mathbf{m}, \mathbf{O}, t)$ with associated weights λ_i for each clique i, $p(\mathbf{m}|\mathbf{O})$ can be defined as a log-linear distribution over the frame indices,

$$p(\mathbf{m}|\mathbf{O}) = \frac{\exp \sum_t \sum_i \lambda_i f_i(\mathbf{m}, \mathbf{O}, t)}{Z(\mathbf{O})}. \quad (4)$$

$Z(\mathbf{O}) = \sum_m \left(\exp \sum_t \sum_i \lambda_i f_i(\mathbf{m}, \mathbf{O}, t) \right)$ is a normalizing constant summing over all possible state sequences $\mathbf{m}$ to form a valid probability distribution.

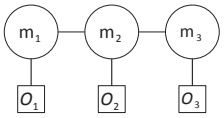

Fig. 2. A graph of linear-chain CRF. Pairs of states (m_1, m_2); (m_2, m_3); and every pair of state and observation vertices $\{(m_t, O_t)\}, t = \{1, 2, 3\}$ form the cliques.

In training process, the optimal weights are obtained by minimizing the negative conditional log-likelihood $-\log p(\mathbf{m}|\mathbf{O})$ over a set of training data X with R sequences, $X = \{(\mathbf{m}_r, \mathbf{O}_r)\}_r^R$.

$$\mathcal{L}_\lambda = \sum_r \left[\left(-\sum_t \sum_i \lambda_i f_i(\mathbf{m}_r, \mathbf{O}_r, t) \right) + \log Z(\mathbf{O}_r) \right] + c\|\lambda\|_2^2, \quad (5)$$

$c\|\lambda\|_2^2$ is a ℓ_2-norm regularization term added to avoid over-training. The level of regularization is controlled by the constant c. The objective function is convex. Globally optimal solutions can be obtained by gradient descent methods for large-scale problems.

In decoding process, the optimal state sequence $\mathbf{m}^*$ maximizes the conditional log-likelihood with the trained λ_i, $\mathbf{m}^* = \arg\max_{\mathbf{m}} \sum_t \sum_i \lambda_i f_i(\mathbf{m}, \mathbf{O}, t)$, by ignoring the constant term $Z(\mathbf{O})$. It can be found by the Viterbi algorithm.

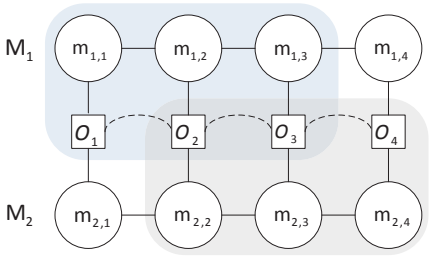

Fig. 3. Single-microphone separation problem expressed as two independent linear-chain CRFs sharing the same observations.

3.2. Modeling speech separation problem with CRF

The two-source separation problem is formulated with the CRF framework. The sources are represented by independent Markov processes $\mathbf{M_k}$ as in factorial HMM, but instead of the generative approach, a direct model approach is applied as follows,

$$(\mathbf{m}_1^*, \mathbf{m}_2^*) = \arg \max_{\mathbf{m}_1, \mathbf{m}_2} p(\mathbf{m}_1, \mathbf{m}_2|\mathbf{O}) \quad (6)$$

We have $p(\mathbf{m}_1, \mathbf{m}_2|\mathbf{O}) = p(\mathbf{m}_1|\mathbf{O})p(\mathbf{m}_2|\mathbf{O})$ given the statistical independence assumption of sources. The separation problem becomes two training and decoding problems for two independent CRFs, each corresponding to one source. The two sources are associated with the same observations $\mathbf{O}$ from the speech mixture.

The problem is illustrated by the graph in Figure 3. For each source k, the state sequence $\mathbf{m_k}$ is associated with vertices $\{m_{k,t}\}$. The vertices $\{O_t\}$ represent observations from the speech mixture. There are edges between $\{O_t\}$ and $\{m_{k,t}\}$, $\{m_{k,t-1}\}$ and $\{m_{k,t}\}$, showing that they are dependent to each other. There is no edge between $\{m_{1,t}\}$ and $\{m_{2,t}\}$, indicating the independence assumption. Dotted edges at observation vertices $\{O_t\}$ indicate they are either dependent or independent of each other. The graph can be split into two linear-chain CRF subgraphs, which share the same observation vertices $\{O_t\}$, as indicated in the shaded region.

A Gaussian distribution set $\mathfrak{G}_k = \{\mathcal{N}(\mu_{k,g}, \Sigma_{k,g}) : g \in G\}$ with G members, is defined to model $p(\hat{s}_{k,t}|m_{k,t})$, the emission probability densities of log power spectra of source signals $\hat{s}_{k,t}$ for each k. Source label sets $\{g : \mathcal{N}(\mu_{k,g}, \Sigma_{k,g}) \in \mathfrak{G}_k\}$ are defined to represent the states $m_{k,t}$ with the corresponding emission probability densities of $\hat{s}_{k,t}$, in a state sequence $\mathbf{m_k}$ at frame t. The source labels are the output of the CRF separation framework.

4. FEATURES FOR CRF SEPARATION FRAMEWORK

4.1. Observations of CRF models

An attribute is a variable describing a property of the speech mixture. Without the conditional independence assumption on observations, an observation modeled by CRF can be an attribute of current frame or a concatenation of attributes from the previous frames up to history length $t-n$. This is especially useful in modeling speech signals for speech separation. Observations from previous frames may provide hints for estimating the sources in the current frame. This type of frame dependency is difficult to be modeled with factorial HMM.

Table 1 shows different types of observations being considered in this paper. Log power spectra of speech mixture are modeled with

Table 1. Observations of CRF framework for single-microphone speech separation problem of two sources.

Category	Description	Attribute	ID	Hist. up to
Log power spectra of mixture	Log power spectra of mixture quantized to one of 2048 clusters of multivariate Gaussian distribution	Cluster index, from 1 to 2048	LS2048	$t-4$
Outputs of factorial HMM	Log power spectra of sources represented by output states estimated by factorial HMM from speech mixture	Source labels of output states, from 1 to $G = \{16, 128, 512\}$	GMMOut	$t-4$
Speaker log-likelihood difference	Computed from speech mixture $\hat{y}_t$ by speaker dependent GMM	clipped at $[-100, 100]$, with interval of 10	SPKR	$t-0$
Fundamental frequency of mixture	Fundamental frequency observed in speech mixture, if available	50 Hz-1 kHz, with 5 Hz interval, 0 if unobserved	F0	$t-1$

concatenation of attributes from previous n frames, in contrast to factorial HMM, which the observation $\hat{y}_t$ is only from current frame. For the computation efficiency, the log power spectra are quantized into one of Gaussian distribution clusters described in Section 5.1 and the history length $n = 4$ is chosen.

The optimal state sequences from factorial HMM are considered as indirect observations of the mixture. By incorporating the results from other separation processes, the CRF framework is implemented as a fusion system. Observations are formed by concatenating the indices associated with the output states up to previous 4 frames.

Human listeners rely on speaker identity for attending to target speaker in multi-talker condition. Speaker information can be provided by measuring speaker log-likelihood. If the frame is dominated by both speakers, the log-likelihood difference is close to zero. For the frames dominated by single speaker, the positivity or negativity of the log-likelihood difference indicates the speaker identity.

Fundamental frequency of the speech mixture is extracted if it is observed. It provides complementary acoustic information of the speakers. Many frames in the mixture are dominated by only one speaker. The observed fundamental frequency is the same as the fundamental frequency of the corresponding speaker. If the frame contains more than one speaker, fundamental frequency is either unobserved or out of the normal ranges of the speakers.

4.2. Feature functions

Feature functions are used to integrate observations into CRF models. Two types of feature functions are defined: the state feature function and the transition feature function. The state feature function describes the relationship between the current observations and the source label. For example, consider a training data at frame t: the j^{th} observation in the speech mixture is fundamental frequency of 150 Hz and the label of source 1 is 10.

$$f_i(\mathbf{m}, \mathbf{O}, t) = \begin{cases} 1, & \text{if } O_t^j = 150 \text{ and } m_{1,t} = 10 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

The transition feature function describes the relationship between two adjacent source states. Following the previous example, the label of source 1 at frame $t-1$ is 9.

$$f_{i+1}(\mathbf{m}, \mathbf{O}, t) = \begin{cases} 1, & \text{if } m_{1,t-1} = 9 \text{ and } m_{1,t} = 10 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

All observations are discrete in value and all feature functions in this paper are binary-valued at $\{0, 1\}$ although these are not requirements of CRF. A feature function can be considered as a count on the observed feature. Discrete observations facilitate this counting process. $\sum_t f_i(\mathbf{m}, \mathbf{O}, t)$ in Equation 4 is equivalent to occurrence frequency. The training process thus depends on sufficient statistics.

5. EXPERIMENTS

5.1. Experimental setup

Speech materials from 3 male and 3 female speakers from the GRID Corpus are selected for the experiments [7]. There are 500 unique utterances for each speaker. The speech materials are mixed into 3 sets of speech mixtures, at SNR[1] of 0 dB. In each set, there are around 2000 speech mixtures mixed from 450 utterances for training. Another 2500 speech mixtures are generated from the remaining 50 utterances for evaluation. Details of speech mixture sets are given in Table 2. The speech materials are re-sampled at 16 kHz. Short-time analysis is applied with Hamming window of 32 ms and frame shift of 10 ms. 257-dimension log-power spectrum is obtained after 512-point fast Fourier transform and logarithm on the power spectrum of each frame. Speaker-dependent GMM are trained from log power spectra of the same 450 utterances with 16, 128 and 512 GMM mixtures for each speaker.

Table 2. Configuration and speaker ID of 3 sets of speech mixtures.

	Male+Male	Male+Female	Female+Female
Speaker 1	1 (Male)	17 (Male)	24 (Female)
Speaker 2	2 (Male)	18 (Female)	25 (Female)

For a GMM with high dimensionality, the training process of GMM can be considered as a clustering process, since most probability mass is concentrated on one Gaussian distribution when it is applied to a sample [8] . The dimensionality of a speaker-dependent GMM is high enough to be viewed as clusters of G multivariate Gaussian distribution if the mixture weights are ignored. The log power spectra of each set of speech mixtures are clustered in the similar manner, for the observations listed in Table 1.

Factorial HMM decoding are performed on the speech mixtures of each set. Transition probabilities between the states are uniform. For CRF training, observations and source labels are extracted from the training samples of the speech mixture sets. The training is performed by CRFsuite software with stochastic sub-gradient method [9]. The same software is used for decoding. After CRF or factorial HMM decoding, log power spectra of sources $\hat{\mathbf{s}}_\mathbf{k}$ are recovered by soft-mask filtering [10]. Waveforms are reconstructed with the mixture phase spectrum by overlap-add method.

5.2. Results and discussion

The evaluation is performed with different settings of observations. They are listed in Table 3. The recovered signals are compared with the original signals from the corpus. Signal-to-noise ratio (SNR), segmental SNR (SSNR), and perceptual evaluation of speech quality (PESQ) are the performance metrics [11]. SNR and SSNR compare

[1]Without loss of generality, the utterance from another speaker is undesired, hence it is noise.

the quality in signal level. PESQ is an objective evaluation metric for human perceived quality. Results of 16, 128, 512 source labels are listed in Table 4. They are the average of two sources within the same mixture set. Result breakdown is shown for the case of 128 source labels, which represents a medium problem size. Key results are shown for 16 and 512 source labels, which represent a small and a large problem size respectively. Oracle results are also included as a reference of the performance limit, in which the source labels are directly extracted from the original sources. A note is that the sources recovered from oracle results are still distorted due to the use of mixture phases. SNR and SSNR may not be always better.

Table 3. Observation setting for CRF separation framework.

	Name	Observation sets
1.	FHMM	Results by factorial HMM (baseline)
2.	CRF-noGMM	SPKR+F0+LS2048
3.	CRF-GMM	GMMOut (GMM retrained by CRF)
4.	CRF-cfgA	GMMOut+F0
5.	CRF-cfgB	GMMOut+SPKR
6.	CRF-cfgC	GMMOut+LS2048
7.	CRF-cfgD	GMMOut+SPKR+F0+LS2048
8.	Oracle	Labels extracted from sources

The results of CRF separation framework consistently improve in terms of PESQ after integrating the observations with the initial separation results from factorial HMM, and usually improve in terms of SNR and SSNR. When a new observation is integrated into the framework, the performance tends to improve, although the improvement may not be dramatic. The results show that log power spectrum, fundamental frequency and speaker likelihoods of the mixture are generally useful. Integrating all these observations further improves the separation results. However, when the problem size continues to grow, the advantage over factorial HMM is smaller.

The CRF separation framework is also evaluated without the initial separation results (**GMMout**) as **CRF-noGMM**. When the problem size is small, such as only 16 source labels, the results are better than that of factorial HMM. When the problem size increases to 128 source labels, the results are still better, except for Male+Male set. This suggests that if there are more different types of observations, the CRF separation framework is competitive to other separation algorithms without integrating the initial results from them.

When the CRF separation framework is based on only a few types of observations, over-training is observed. An example is the Male+Male set with only **GMMout** as observation (**CRF-GMM**). The results are significantly poorer. When more types of observations are included, the improvement trend is restored.

6. CONCLUSION

In this paper, a CRF based framework for single-microphone separation is proposed. The framework is able to integrate different types of observations and even incorporate the results of other separation processes. The observations are not required to be conditional independent of each other, hence a large variety of observations can be chosen. Experiments show that the separation results from factorial HMM separation approach are improved by the proposed method when log power spectrum, fundamental frequency and speaker likelihoods of the mixture are integrated. Without the initial separation results, the framework can reach or over the performance level of factorial HMM with 16 and 128 output source labels. It is believed that the performance can be further improved if more different types of observations are available. Currently, the framework is based on speaker-dependent models. Our future work is to extend the framework to be speaker independent.

Table 4. Results in SNR (dB), SSNR (dB) and PESQ of CRF separation framework for $G = \{16, 128, 512\}$ source labels.

Mix.		Male + Male			Male + Female			Female + Female		
		SNR	SSNR	PESQ	SNR	SSNR	PESQ	SNR	SSNR	PESQ
16	FHMM	3.49	1.83	1.63	4.61	3.37	1.68	3.74	3.10	1.46
	CRF-noGMM	3.74	2.46	1.78	5.19	3.72	1.92	3.82	3.45	1.64
	CRF-GMM	3.43	2.19	1.67	4.97	3.59	1.84	3.67	3.20	1.53
	CRF-cfgD	***3.79***	***2.49***	***1.81***	***5.21***	***3.77***	***1.94***	***3.86***	***3.51***	***1.66***
	Oracle	4.37	3.18	2.15	4.11	3.01	2.06	5.14	6.07	2.19
128	FHMM	4.16	2.72	1.87	6.10	4.49	2.06	5.15	4.16	1.82
	CRF-noGMM	3.80	2.68	1.78	6.16	4.55	2.08	4.99	4.31	1.89
	CRF-GMM	3.85	2.77	1.71	6.17	4.64	2.05	5.13	4.43	1.89
	CRF-cfgA	3.85	2.78	1.71	6.27	4.68	2.10	5.23	4.51	1.93
	CRF-cfgB	4.04	2.86	1.82	6.28	4.69	2.11	5.21	4.49	1.95
	CRF-cfgC	4.05	2.87	1.82	6.29	4.69	2.11	5.19	4.50	1.92
	CRF-cfgD	***4.22***	***2.94***	***1.93***	***6.40***	***4.75***	***2.18***	***5.28***	***4.58***	***1.98***
	Oracle	5.21	3.86	2.51	5.22	3.91	2.39	5.90	5.18	2.35
512	FHMM	4.16	2.79	1.96	6.72	4.68	2.24	***5.86***	4.64	2.01
	CRF-noGMM	3.94	2.69	1.90	6.49	4.64	2.15	5.44	4.55	1.97
	CRF-GMM	4.20	2.97	1.93	6.74	4.83	2.20	5.79	4.83	2.08
	CRF-cfgD	***4.30***	***2.98***	***2.01***	***6.85***	***4.89***	***2.27***	5.81	***4.84***	***2.11***
	Oracle	5.55	4.13	2.66	5.75	5.62	2.89	6.73	5.60	2.58

7. REFERENCES

[1] Martin Cooke, John R Hershey, and Steven J Rennie, "Monaural speech separation and recognition challenge," *Comp., Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[2] Sam T. Roweis, "One microphone source separation," in *In Advances in Neural Information Processing Systems 13*. 2001, pp. 793–799, MIT Press.

[3] J D Lafferty, A McCallum, and F Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proc. ICML*, 2001, pp. 282–289.

[4] A Nadas, D Nahamoo, and M A Picheny, "Speech recognition using noise-adaptive prototypes," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 10, pp. 1495–1503, 1989.

[5] M H Radfar, A H Banihashemi, R M Dansereau, and A Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, pp. 724, 2006.

[6] Julian Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974.

[7] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[8] N Merhav and Y Ephraim, "Hidden Markov modeling using a dominant state sequence with application to speech recognition," *Comp., Speech and Lang.*, vol. 5, no. 4, pp. 327–339, 1991.

[9] Naoaki Okazaki, "CRFsuite: a fast implementation of Conditional Random Fields (CRFs)," 2007.

[10] M H Radfar and R M Dansereau, "Single-Channel Speech Separation Using Soft Mask Filtering," *IEEE Trans. Audio, Speech, and Lang. Prcs.*, vol. 15, no. 8, pp. 2299–2310, 2007.

[11] Y Hu and P C Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. Audio, Speech, and Lang. Prcs.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.