MODEL-BASED INDEPENDENT COMPONENT ANALYSIS FOR ROBUST MULTI-MICROPHONE AUTOMATIC SPEECH RECOGNITION

Laurent Couvreur *Christophe Ris*

Faculté Polytechnique de Mons
Avenue Copernic 1
7000 Mons, Belgium
lcouv,ris@tcts.fpms.ac.be

ABSTRACT

In this communication, we present a method for noise-robust multi-microphone automatic speech recognition (ASR). It is assumed that the speech source to be recognized is recorded with several microphones in a noisy acoustic environment. The proposed method estimates the short-term subband energies (as they are needed for computing the ASR front-end) of the clean speech source from the ones of the microphone noisy signals. The estimation procedure is based on the concept of Independent Component Analysis (ICA) and it is driven by the acoustic model used by the ASR decoder. The method is shown to be highly robust for a connected digit recognition task in high noise conditions, improving word error rates by more than 50% relatively to the performance of the baseline single-microphone ASR system.

1. INTRODUCTION

Automatic speech recognition (ASR) is a key component for hands-free man-machine interaction. State-of-the-art ASR systems are based on statistical acoustic models which are commonly trained on clean material, i.e. noise-free speech. In many applications (e.g., mall directory assistance, automatic ticketing machine, etc), ASR systems are operated in noisy environments. Consequently, their performances degrade severely because of the mismatch between the training conditions (clean speech) and the operating conditions (noisy speech).

A typical phoneme-based ASR system, as it is considered in this work, is depicted in figure 1. It consists of three main blocks. First, the front-end (FE) chops the speech signal recorded at the microphone into frames and computes a set of relevant acoustic coefficients for every frame. Next, acoustic coefficient vectors are fed into the acoustic model (MA) which estimates a probability score for every phoneme. Finally, the word decoder (DEC) searches for the most likely word string, under the constraint of some phonetic lexicon and word grammar, given the sequence of phoneme probability vectors for all the frames. The FE algorithm typically consists in computing cepstral coefficients which are derived from the frames energies obtained in some frequency subbands [1]. In noisy environments, the recorded speech signal is corrupted by additive noise, and so are the subband energies and the cepstral acoustic coefficients. Hence, the acoustic model produces unreliable probability scores and the decoding search is misled to incorrect recognition results.

In this work, we propose to apply a technique based on Independent Component Analysis (ICA) [2] to estimate the subband energies of clean speech frames from the subband energies of noisy speech frames recorded at several microphones within a noisy operating environment. Unlike classical ICA techniques, our approach consists in extracting only a single independent component (IC) from the observed mixtures. To guarantee that the single IC corresponds to the clean speech subband energies, the IC search is steered by the acoustic model used by the ASR decoder. That is, the IC is searched in order to match the acoustic model as well as possible by minimizing the mean frame negative entropy of the phoneme probability scores.

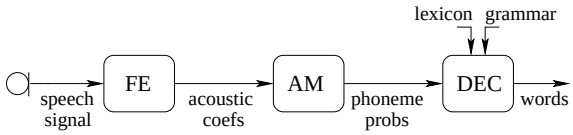

Fig. 1. A typical phoneme-based ASR system: microphone, front-end (FE), acoustic model (MA) and word decoder (DEC).

In the next section, we first describe the signal model which is assumed in this work and some notations are introduced. Then, we formulate the problem of estimating the subband energies of clean speech recorded via several microphones in a noisy environment in terms of ICA processing. In section 3, the ICA-based estimation algorithm is presented. The method is applied for recognition of connected digit sequences in noisy conditions. Results demonstrating the efficiency of the proposed method are reported in section 4. Conclusions are drawn in section 5.

2. SIGNAL MODEL

2.1. Single-Microphone Model

Consider that the speech source s_n to be recognized is embedded in a noisy environment. That is, the speech source coexists with several noise sources which can be non-speech sound sources or minor speech sources. If we assume that the operating environment is acoustically open (i.e., low reverberation) and that the sources are fixed, then the signal x_n recorded at a distant single microphone is given by

$$x_n = \alpha_s s_{n-n_s} + \sum_{i=1}^{R} \alpha_i v_{i,n-n_i} \qquad (1)$$

where $v_{i,n}$ denotes the ith competitive noise source. The attenuation factors $\alpha_{(s,i)}$ and the delays $n_{(s,i)}$ are solely defined by the source locations with respect to the microphone. Note that the number R of noise sources is generally unknown. In our baseline

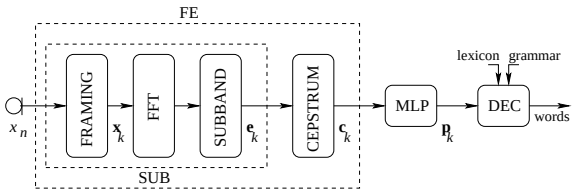

Fig. 2. Single-microphone ASR system: subband analysis (SUB), cepstrum transform, acoustic model (MLP) and word decoder (DEC).

single-microphone ASR system (see figure 2), the microphone signal x_n is first divided into K overlapping frames. Define[1] $\mathbf{x}_k = [x_{kN_r}, \ldots, x_{kN_r+N_w-1}]^T$ as the kth frame, $1 \leq k \leq K$, with $N_r = F_s/F_r$ and $N_w = T_w \times F_s$, where F_s, F_r and T_w denote the sampling frequency [Hz], the frame rate [Hz] and the frame length [s], respectively. The frame $\mathbf{x}_k$ is then weighted by the Hamming window and its short-term frequency spectrum is computed by Fast Fourier Transform (FFT). Next, subband energies are computed by applying a filterbank and integrating the power spectrum into every frequency subband. In this work, the power spectrum is simply weighted by triangular non-uniform MEL filters as described in [1] and integrated to form the subband energy vector $\mathbf{e}_k = [e_k^1, \ldots, e_k^L]^T$ of the kth frame with L denoting the number of frequency subbands. As an example, subband energies for a typical clean speech utterance is given in figure 5(c). Finally, a cepstral vector $\mathbf{c}_k = [c_k^1, \ldots, c_k^P]^T$, with P being the number of cepstral coefficients, is computed by taking the logarithm of $\mathbf{e}_k$ and applying an Inverse Discrete Cosine Transform (IDCT). The cepstral vectors are then fed into the acoustic model in order to estimate the phoneme probability vector $\mathbf{p}_k = [p_k^1, \ldots, p_k^Q]^T$ with Q standing for the number of phonemes. In this work, we use a hybrid Multi Layer Perceptron (MLP) / Hidden Markov Models (HMM) ASR system [3]. That is, the *a posteriori* probability $p_k^q = P(\text{phoneme } q|\mathbf{c}_k)$ that the qth phoneme is uttered while observing $\mathbf{c}_k$ is estimated with a MLP. In noisy environments, the recorded speech signal is corrupted by additive noise, and so are the subband energies and the cepstral coefficients. Hence, the acoustic model which has be trained on clean speech cepstral vectors produces unreliable probability vectors and the decoding search is misled to incorrect recognition results.

2.2. Multi-Microphone Model

Next, we consider that the speech source within a noisy environment is recorded with M ($M > 1$) distant microphones. For every microphone, we can write

$$x_{j,n} = \alpha_{sj} s_{n-n_{sj}} + \sum_{i=1}^{R} \alpha_{ij} v_{i,n-n_{ij}}, \quad 1 \leq j \leq M, \qquad (2)$$

with $\alpha_{(s,i)j}$ and $n_{(s,i)j}$ being the attenuation factors and the delays between the (s,i)th source and the jth microphone. Under the assumption that the sources are mutually independent in a statistical sense, and that the attenuation factors do not depend on the frequency, equation (2) can be written in the subband energy domain as follows

$$\mathbf{e}_{x_j,k} = \alpha_{sj}^2 \mathbf{e}_{s,k-k_{sj}} + \sum_{i=1}^{R} \alpha_{ij}^2 \mathbf{e}_{v_i,k-k_{ij}}, \quad 1 \leq j \leq M, \quad (3)$$

[1]The notation $\mathbf{x}^T$ denotes the transpose of vector $\mathbf{x}$.

where $\mathbf{e}_{x_j,k}$, $\mathbf{e}_{s,k}$ and $\mathbf{e}_{v_i,k}$ stand for the sequences of subband energy vectors of the jth microphone signal, the target speech source and the noise sources, respectively. Note that equation (3) is not exactly valid, yet it is a fair approximation [4]. Besides, a typical frame rate F_r is 100Hz which corresponds to a frame shift of 10ms. For a standard 342m/s sound speed, it suggests that the propagation delays in the subband energy domain for a given source are roughly equal unless the micorphones are more distant from each other than about 3m. Hence, if we assume that the microphones are close enough, equation (3) can be approximated by

$$\mathbf{e}_{x_j,k} \approx \alpha_{sj}^2 \mathbf{e}_{s,k-k_s} + \sum_{i=1}^{R} \alpha_{ij}^2 \mathbf{e}_{v_i,k-k_i}, \quad 1 \leq j \leq M, \qquad (4)$$

where the delays are constant for all the microphone locations. Equivalently, equation (4) can be written in a scalar form for every frequency subband,

$$e_{x_j,k}^l \approx \alpha_{sj}^2 e_{s,k-k_s}^l + \sum_{i=1}^{R} \alpha_{ij}^2 e_{v_i,k-k_i}^l, \quad 1 \leq l \leq L. \qquad (5)$$

That is, we have formulated the model of recording a speech source in a noisy environment as a linear instantaneous mixing model in the subband energy domain,

$$\begin{bmatrix} e_{x_1,k}^l \\ \vdots \\ e_{x_M,k}^l \end{bmatrix} = \begin{bmatrix} \alpha_{s1}^2 & \alpha_{11}^2 & \ldots & \alpha_{R1}^2 \\ \vdots & \vdots & & \vdots \\ \alpha_{sM}^2 & \alpha_{1M}^2 & \ldots & \alpha_{RM}^2 \end{bmatrix} \begin{bmatrix} e_{s,k}^l \\ e_{v_1,k}^l \\ \vdots \\ e_{v_R,k}^l \end{bmatrix} \qquad (6)$$

for $1 \leq k \leq K$ and $1 \leq l \leq L$. One can write equation (6) in a compact form,

$$\mathbf{e}_x = \mathbf{A} \begin{bmatrix} e_s \\ \mathbf{e}_v \end{bmatrix} \qquad (7)$$

where the mixed components $[e_s \ \mathbf{e}_v^T]^T$ and the mixing matrix $\mathbf{A}$ are unknown. In the following, we present an algorithm for estimating the mixed component e_s, i.e. the subband energies $e_{s,k}^l$, from the observed mixtures $\mathbf{e}_x$, i.e. the subband energies $e_{x_j,k}^l$. The algorithm aims at restoring clean speech subband energy vectors in order to compute cepstral vectors as they are expected by the acoustic model and thus estimate reliably phoneme probability vectors.

3. PROPOSED METHOD

In this work, we propose to apply a technique based on Independent Component Analysis (ICA) [2] to problem (7) for estimating the subband energies $e_{s,k}^l$ of clean speech frames from the subband energies $e_{x_j,k}^l$ of noisy speech frames recorded at several microphones within a noisy operating environment. Under the assumptions that at most one mixed component is Gaussian and that the mixing matrix $\mathbf{A}$ is full column rank, the problem is identifiable [2] and ICA of any random vector $\mathbf{e}_x$ is defined as a linear transformation

$$\begin{bmatrix} \hat{e}_s \\ \hat{\mathbf{e}}_v \end{bmatrix} = \mathbf{W} \ \mathbf{e}_x \qquad (8)$$

where the unmixing matrix $\mathbf{W}$ is determined so that the components of the transformed vector, the so-called Independent Components (IC), are statistically as independent as possible. Adopting an information-theoretic approach, it is equivalent to search $\mathbf{W}$ in

order to minimize the mutual information of the transformed components [2]. Actually, complete ICA of $\mathbf{e}_x$ is superfluous: we are interested in only the speech-related IC. Besides, the number of mixed components is generally unknown and IC's can be obtained only up to a permutation. Hence, we prefer adopting a deflation approach as described in [5] and find only one IC,

$$\hat{e}_s = \mathbf{w}^T \mathbf{e}_x \tag{9}$$

where the unmixing vector $\mathbf{w}$ is determined so that the negentropy $J(\hat{e}_s)$ of the extracted IC is maximized [2]. In the framework of deflation-based ICA, the negentropy can be approximated by the so-called contrast function $J_G(\mathbf{w})$ [5],

$$J(\hat{e}_s) \approx J_G(\mathbf{w}) = (E\{G(\mathbf{w}^T\mathbf{e}_x)\} - E\{G(\nu)\})^2 \tag{10}$$

where $E\{\}$ denotes the expectation operator, $G()$ is a smooth non-quadratic function and ν is a Gaussian random variable with zero mean and unit variance. That is, we have to solve the following maximization problem,

$$\mathbf{w} = \arg\max_{\mathbf{w}} J_G(\mathbf{w}). \tag{11}$$

Note that the observed vectors $\mathbf{e}_x$ are centered and whitened to make the estimation process simpler to derive and better conditioned [5]. Dewhitening and decentering are performed afterwards. In order to be meaningful, the maximization has to be performed under an additional constraint, e.g. $E\{(\mathbf{w}^T\mathbf{e}_x)^2\} = 1$. Since, $\mathbf{e}_x$ has been whitened, the constraint reduces to $h(\mathbf{w}) = \|w\|^2 - 1 = 0$. The maxima of $J_G(\mathbf{w})$ are obtained at some optima of $E\{G(\mathbf{w}^T\mathbf{e}_x)\}$. Hence, we have to solve an optimization problem under a single equality constraint. According to the Kuhn-Tucker conditions [6], the optima are obtained at points where

$$\begin{aligned} \mathbf{f}(\mathbf{w}) &= \nabla E\{G(\mathbf{w}^T\mathbf{e}_x)\} - \lambda\nabla h(\mathbf{w}) \\ &= E\{\mathbf{e}_x g(\mathbf{w}^T\mathbf{e}_x) - \lambda\mathbf{w} = 0 \end{aligned} \tag{12}$$

with $g()$ being the first derivative of $G()$. The Lagrange multiplier λ is simply equal to $E\{\mathbf{w}^T\mathbf{e}_x g(\mathbf{w}^T\mathbf{e}_x)\}$. Equation (12) can be solved for $\mathbf{w}$ by an iterative Newton's method [6],

$$\mathbf{w}^{(\ell+1)} = \mathbf{w}^{(\ell)} - [\mathbf{F}(\mathbf{w}^{(\ell)})]^{-1}\mathbf{f}(\mathbf{w}^{(\ell)}) \tag{13}$$

where $\mathbf{F}$ is the Jacobian matrix of the vector function $\mathbf{f}$. The vector function $\mathbf{f}$ can be computed analytically for a given function G, while the Jacobian matrix $\mathbf{F}$ is generally computed numerically. In this work, we use the general-purpose G function given in [5],

$$G(y) = 1/a_1 \log\cosh(a_1 y), \quad g(y) = \tanh(a_1 y) \tag{14}$$

with $a_1 \approx 1$. The expectations are computed for all the data in a batch mode. The initial unmixing vector $\mathbf{w}^{(0)}$ is chosen randomly. A major problem is to guarantee that the single extracted IC is the one relative to the speech source. To do so, we modify equation (12) for attracting the search path to a zero point such that the resulting IC matches the acoustic model of the ASR system as well as possible. A natural way of characterizing the acoustic model adequacy is the mean frame negative entropy H

$$H = -\frac{1}{K}\sum_{k=1}^{K}\sum_{q=1}^{Q} p_k^q \log p_k^q \tag{15}$$

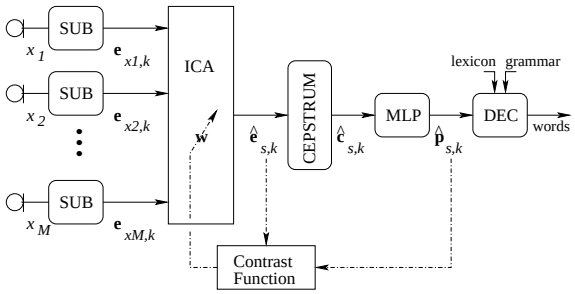

Fig. 3. ICA-based multi-microphone ASR system with acoustic model feedback.

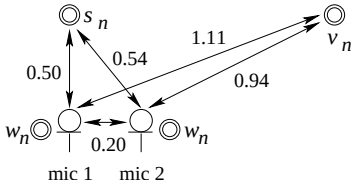

Fig. 4. Simulated recording scenario for ASR experiments: speech source s_n, located noise source v_n and diffuse noise source w_n. Distances are expressed in meters.

which is always non-negative. The better the input data match the acoustic model, the lower the entropy is (ideally equal to zero). That is, the modified vector function $\mathbf{f}$ is defined as

$$\mathbf{f}(\mathbf{w}) = E\{\mathbf{e}_x g(\mathbf{w}^T\mathbf{e}_x) + \nabla H(\mathbf{w}) - \lambda\mathbf{w} = 0 \tag{16}$$

where the gradient $\nabla H(\mathbf{w})$ is evaluated numerically. The resulting multi-microphone ASR system is depicted in figure 3.

4. EXPERIMENTAL RESULTS

Consider the simple scenario described in figure 4. A speech source s_n that we want to recognize is recorded with two distant microphones in the presence of a located noise source v_n,

$$x_{1,n} = \alpha_{s1}s_{n-n_{s1}} + \alpha_{v1}v_{n-n_{v1}} + w_n \tag{17}$$

$$x_{2,n} = \alpha_{s2}s_{n-n_{s2}} + \alpha_{v2}v_{n-n_{v2}} + w_n \tag{18}$$

where w_n models some background diffuse noise (i.e., same contribution at both microphones). The delays and the attenuation factors are set proportional and inversely proportional, respectively, to the distances (see figure 4) between the sources and the microphones. The relative amplitude of the noise sources with respect to the speech source is controlled by the signal-to-noise ratios $SNR_v = 10\log_{10} E\{s_n^2\}/E\{v_n^2\}$ and $SNR_w = 10\log_{10} E\{s_n^2\}\ /E\{w_n^2\}$. In our experiments, the speech and noise material come from the AURORA database [7]. The training and testing speech databases consist of English connected digit sequences. The located and diffuse noise sources are of "babble" and "airport" types, respectively. The noise signals change for each utterance. A typical digit sequence and its noisy version are given in figure 5.

First, we evaluate our baseline single-microphone ASR system (see figure 2). The FE algorithm is applied only to the signal measured at the left microphone with $F_s = 8000$Hz, $T_w = 30$ms, $F_r = 100$Hz, the number of subbands $L = 24$ and the number of cepstral coefficients $P = 12$. The acoustic model is a 600-node

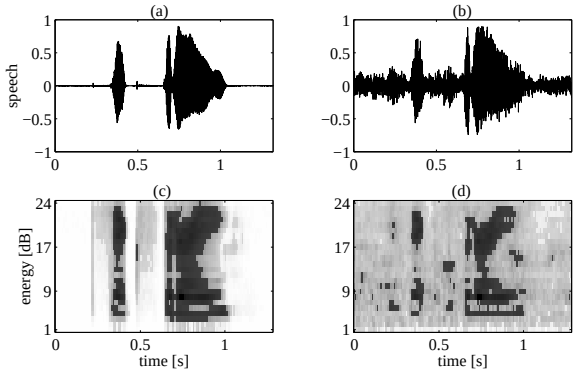

Fig. 5. Waveform and subband energies of (a)-(c) clean speech utterance "63" and (b)-(d) its noisy versions for $SNR_v = 5$dB and $SNR_w = 15$dB.

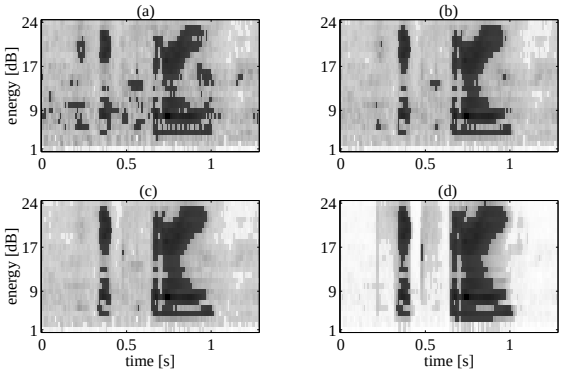

Fig. 6. Subband energies of extracted IC after (a) 1 iteration, (b) 2 iterations, (c) 5 iterations and (d) after convergence for noisy speech utterance "63" with $SNR_v = 5$dB and $SNR_w = 15$dB.

Table 1. Word error rates [%] of baseline single-microphone ASR.

SNR_v	SNR_w						
	∞	20dB	15dB	10dB	5dB	0dB	-5dB
∞	**1.4**	5.0	10.9	22.5	43.2	69.2	90.3
20dB	5.1	7.9	13.3	24.1	43.0	69.7	89.2
15dB	10.6	13.2	17.3	28.0	45.3	70.7	90.1
10dB	23.7	24.9	27.7	35.8	51.6	73.6	90.9
5dB	45.6	47.5	49.3	54.2	63.9	79.2	92.3
0dB	73.8	74.0	75.7	77.7	81.0	88.5	96.0
-5dB	93.7	93.5	93.0	93.6	94.5	96.7	98.7

Table 2. Word error rates [%] of ICA-based 2-microphone ASR.

SNR_v	SNR_w						
	∞	20dB	15dB	10dB	5dB	0dB	-5dB
∞	–	2.3	4.9	6.7	7.1	7.5	7.7
20dB	2.9	4.8	7.1	10.4.	11.7	12.0	12.2
15dB	5.1	8.2	10.2	16.5	18.9	19.2	19.4
10dB	11.8	15.6	18.2	22.8	23.3	23.7	23.6
5dB	15.2	18.3	20.1	23.7	24.5	24.9	25.0
0dB	16.9	19.6	21.4	25.8	26.9	27.4	27.7
-5dB	18.2	20.0	22.0	28.3	28.9	29.2	29.6

single hidden layer MLP which takes nine context vectors to estimate a 33-dimensional phoneme probability vector for each frame, each context vector containing 12 cepstral coefficients and the first derivative of the frame energy. Word decoding is done by Viterbi search with neither pruning nor grammar constraints. Table 1 reports word error rates (WER), defined as the sum of substitution, deletion and insertion error rates, for various noise levels. Clearly, the higher the noise level is, the most severely the performances degrade.

Next, we apply our ICA-based method. First, subband energies are computed for both microphone signals (see figure 3 with $M = 2$) as in the single-microphone FE. Then, the unmixing vector $\mathbf{w}^{(\ell)}$ is estimated. Finally, the subband energies of the clean speech source are extracted and they are fed into the following blocks of the ASR system. Note that the estimation is performed utterance by utterance. The convergence of the ICA-based estimation algorithm is demonstrated in figure 6 for a typical digit sequence. Recognition results are given in table 2. It clearly shows that our ICA-based multi-microphone ASR system outperforms the baseline single-microphone ASR system, especially in high noise conditions.

5. CONCLUSIONS

We have proposed an ICA-based method for estimating the subband energies of a speech signal recorded with several microphones in a noisy environment. Unlike classical ICA techniques, only one IC is extracted from the observed mixtures and the extraction process is directed to the speech-related IC by using a speech statistical model, namely the acoustic model of an ASR system. The method is shown to improve significantly the performance of a multi-microphone ASR system operated in high noise conditions. Future work will consist in extending the method to reverberant conditions (e.g., by working in the energy modulation frequency domain). Besides, the algorithm should be rendered adaptive to allow moving sources.

6. REFERENCES

[1] J. W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1214–1247, Sep. 1993.

[2] P. Common, "Independent Component Analysis: A New Concept ?", *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[3] H. Bourlard and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", *Kluwer Academic Publishers*, 1994.

[4] C. Avendano, "Temporal Processing of Speech in a Time-Feature Space", *Ph.D. Thesis*, Oregon Graduate Institute of Science and Technology, Apr. 1997.

[5] A. Hyvärynen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, Mar. 1999.

[6] D. G. Luenberger, "Linear and Nonlinear Programming", *Addison-Wesley*, 2nd edition, 1984.

[7] AURORA 2.0 – http://www.elda.fr/proj/aurora2.html.