PARTICLE FILTER BASED SOFT-MASK ESTIMATION FOR MISSING FEATURE RECONSTRUCTION

Friedrich Faubel, Humza Raja, John McDonough, Dietrich Klakow

Spoken Language Systems,
Saarland University, D66123 Saarbrücken, Germany
{friedrich.faubel,john.mcdonough,dietrich.klakow}@lsv.uni-saarland.de

ABSTRACT

In this work, we show how *particle filter* (PF) based speech feature enhancement can profitably be combined with soft-decision *missing feature reconstruction*. The combined approach is motivated by the fact that standard minimum mean square error noise compensation techniques fail to give accurate estimates of the clean speech spectrum if the noise spectral power significantly exceeds that of speech in a particular spectral region. Experiments show that the proposed algorithm can reduce the word error rate by up to 26.1% relative, compared to 17.0% for speech feature enhancement based solely on particle filters.

Index Terms— missing feature reconstruction, soft-decision, mean imputation, particle filter, speech feature enhancement

1. INTRODUCTION

Noise compensation methods such as the *vector Taylor series* (VTS) approach [1], sequential *expectation maximization* (EM) [2], *interacting multiple models* (IMMs) [3] or *particle filters* (PFs) [4, 5] typically first form a *minimum mean square error* (MMSE) estimate of the noise that corrupts speech and thereafter compensate for it. This works well as long as the noise spectral power does not significantly exceed that of speech. If it does, the affected portion of the speech spectrum is occluded as portrayed in Figure 1 and it is impossible to say what the underlying clean speech spectrum was. Figure 2 shows how clean speech with a power of 0, 15, 20 and 25 dB respectively is distorted by additive noise using the well-established model [1, 2, 3, 4, 5] of acoustic distortion in the log spectral domain. At an observed power of 35 dB – marked by a dotted horizontal line – the four curves are very close to each other. Thus, a slight misestimation of the noise power can cause the MMSE clean speech estimate to vary between approximately 0 and 30 dB.

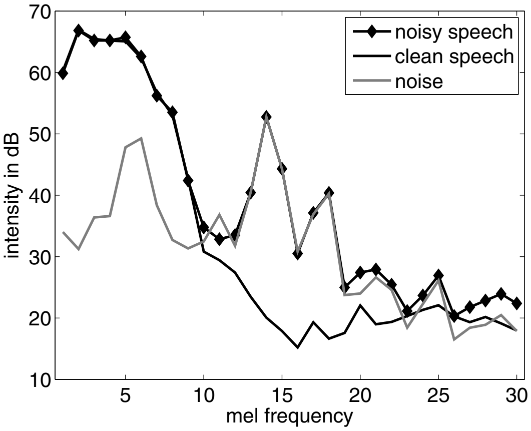

Fig. 1. *Occlusion: in regions where the noise is over 10dB louder than speech the observed noisy speech spectrum is effectively independent of the clean speech spectrum.*

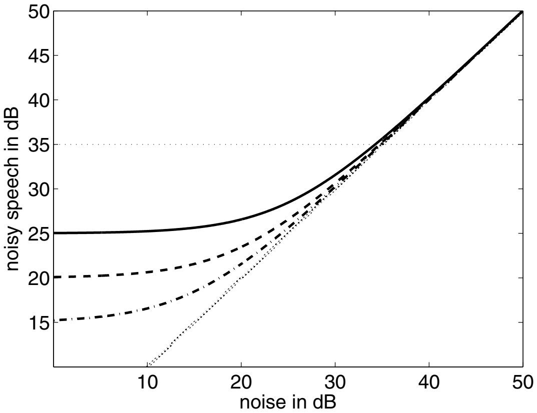

Fig. 2. *Effect of additive noise to a clean speech spectral bin of 25 (solid curve), 20 (dashed curve), 15 (dotted dashed curve) and 0 dB (dotted curve), respectively*

Missing feature reconstruction avoids this problem by explicitly modeling the effect of noise as occlusion of the clean speech spectrum, but on the other hand fails if the noise is not strong enough to actually cause occlusion. Hence, we propose here the combination of MMSE speech feature enhancement and missing feature reconstruction, as originally described in [6] for spectral subtraction. As the noise and therefore the occluded regions of the speech spectrum are likely to vary in time, we use a particle filter to both enhance speech and simultaneously obtain a running estimate for the probability of occlusion. This probability is subsequently used to reconstruct the occluded spectral bins with the soft-decision mean imputation approach explained in Section 2. Section 3 shows how the probability of occlusion can be computed within the particle filter framework.

2. MISSING FEATURE RECONSTRUCTION

Missing feature approaches model the effect of noise in the log Mel domain as occlusion of the clean speech spectrum. The occluded portion is considered to be lost or “missing” and the objective of missing feature reconstruction is to reconstruct this missing portion based on its statistical relationship to the unoccluded portions of the spectrum. If the clean speech distribution is modeled as a Gaussian, the statistical relationship can be the mean and covariance as is in conditional mean imputation [7] or just the mean of the occluded part as in mean imputation [7]. This is usually extended to the use of Gaussian mixtures, as it is well known that the statics of speech are strongly dependent on the phoneme currently being spoken.

In the remaining part of this section we give an alternate and slightly generalized derivation of Raj’s soft-decision mean imputation approach [8], which can be regarded as a refinement of the soft-decision mean imputation method devised in [9]. The approach is based on a diagonal covariance Gaussian mixture model for clean speech and reconstructs the clean speech spectrum as a weighted sum of mean imputations of the individual Gaussians.

2.1. A model for occlusion

Denoting noisy speech, clean speech and noise spectra in the log Mel domain as $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ respectively, the occlusion can be formally expressed as

$$y_d = \max(x_d, n_d), \quad d = 1, \ldots, D, \tag{1}$$

where D is the dimensionality of the Mel filterbank. The occluded components are typically given by a *mask* $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_D\}$ with

$$\theta_d = \begin{cases} 1, & x_d \text{ is observable} \\ 0, & x_d \text{ is occluded} \end{cases} \tag{2}$$

As there might be uncertainty in θ_d, the use of *soft-decision masks* or simply *soft-masks* has been proposed in [9], [10] and [8], whereby the decision in (2) is replaced by the probability of x_d being observable:

$$\theta_d = \mathcal{P}(x_d \geq n_d). \tag{3}$$

When a portion of the clean speech spectrum is occluded, all that is known is that it must lie below the observed noisy speech spectrum. Hence, it is reasonable to assume that an occluded clean speech component x_d is bounded above by y_d. As we add 1 to each of the Mel bins before taking the logarithm – as common in speech recognition – it is also bounded below by 0. Consequently, the distribution of the missing components of the clean speech spectrum can be modeled by truncating the original distribution. As clean speech is typically modeled as a Gaussian mixture

$$p(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k)$$

with diagonal covariance matrices $\boldsymbol{\Sigma}_k = \operatorname{diag}(\sigma_{k,1}^2, \ldots, \sigma_{k,D}^2)$, this devolves to truncating Gaussians.

2.2. The doubly truncated Gaussian distribution

Truncated Gaussian distributions are well known in the statistical literature. In our case, we are dealing with doubly truncated Gaussian distributions [11], which are defined as

$$\mathcal{N}^{[L,U]}(x; \mu, \sigma^2) \triangleq \frac{\mathcal{N}(x; \mu, \sigma^2)\big|_L^U}{\mathcal{C}(U; \mu, \sigma^2) - \mathcal{C}(L; \mu, \sigma^2)}, \tag{4}$$

where $\mathcal{C}$ denotes the cumulative Gaussian distribution, and where $\mathcal{N}(x; \mu, \sigma^2)\big|_L^U$ is $\mathcal{N}(x; \mu, \sigma^2)$ on the interval $[L, U]$, zero outside.

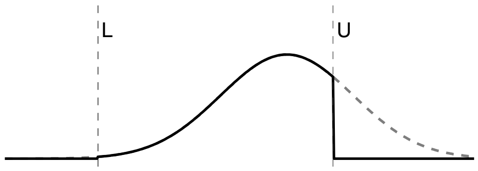

Fig. 3. *Gaussian (dashed grey curve), truncated Gaussian (solid curve)*

Truncating a Gaussian also changes its mean, which was not taken into account in [7, 9], but was in [8]. For a doubly truncated Gaussian the mean can be shown [11] to be

$$\begin{aligned} \mu^{[L,U]} &\triangleq \int_L^U x \mathcal{N}^{[L,U]}(x; \mu, \sigma^2) dx \\ &= \mu - \sigma^2 \frac{\mathcal{N}(U; \mu, \sigma^2) - \mathcal{N}(L; \mu, \sigma^2)}{\mathcal{C}(U; \mu, \sigma^2) - \mathcal{C}(L; \mu, \sigma^2)}, \end{aligned} \tag{5}$$

which will directly be used in the upcoming section.

2.3. Mean imputation

The idea behind mean imputation is to replace the occluded clean speech spectral bins with their means, or, in light of the discussion in the previous section, with their truncated means. For soft-masks this is slightly more complicated. Assuming the clean speech spectrum $\mathbf{x}$ comes from the k-th Gaussian of the mixture, the likelihood of x_d given the observation $\mathbf{y}$ and mask $\boldsymbol{\theta}$ is

$$p(x_d|\mathbf{y}, \boldsymbol{\theta}, k) = \theta_d \delta_{y_d}(x_d) + (1 - \theta_d) \mathcal{N}^{[0,y_d]}(x_d; \mu_{k,d}, \sigma_{k,d}).$$

where δ_{y_d} is a Dirac-delta translated to y_d. Hence, the corresponding mean is an interpolation of the observed noisy speech spectrum with the mean of the truncated Gaussian:

$$\begin{aligned} E\left[x_d|\mathbf{y}, \boldsymbol{\theta}, k\right] &= \int_0^{y_d} x_d p(x_d|\mathbf{y}, \boldsymbol{\theta}, k) dx_d \\ &= \theta_d y_d + (1 - \theta_d) \mu_{k,d}^{[0,y_d]}. \end{aligned} \tag{6}$$

This can be extended to Gaussian mixtures by marginalization over the mixture components,

$$p(x_d|\mathbf{y}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(x_d, k|\mathbf{y}, \boldsymbol{\theta}).$$

Then, rewriting $p(x_d, k|\mathbf{y}, \boldsymbol{\theta})$ as $p(x_d|\mathbf{y}, \boldsymbol{\theta}, k) p(k|\mathbf{y}, \boldsymbol{\theta})$ the mean of the truncated mixture can be shown to be

$$\begin{aligned} E\left[x_d|\mathbf{y}, \boldsymbol{\theta}\right] &= \int_0^{y_d} x_d \sum_{k=1}^{K} p(x_d|\mathbf{y}, \boldsymbol{\theta}, k)\, p(k|\mathbf{y}, \boldsymbol{\theta})\, dx_d \\ &= \sum_{k=1}^{K} p(k|\mathbf{y}, \boldsymbol{\theta}) E\left[x_d|\mathbf{y}, \boldsymbol{\theta}, k\right], \end{aligned} \tag{7}$$

where $p(k|\mathbf{y}, \boldsymbol{\theta})$ is the posterior probability that the underlying clean speech spectrum belongs to the k-th Gaussian.

2.4. Posterior probability of a particular Gaussian

Using Bayes' rule the probability that clean speech originated from the k-th Gaussian, given the observation $\mathbf{y}$ and soft-mask $\boldsymbol{\theta}$, can be expressed as

$$p(k|\mathbf{y},\boldsymbol{\theta}) = \frac{p(\mathbf{y}|\boldsymbol{\theta},k)p(k)}{p(\mathbf{y}|\boldsymbol{\theta})} = \frac{c_k p(\mathbf{y}|\boldsymbol{\theta},k)}{\sum_{k'=1}^{K} c_{k'} p(\mathbf{y}|\boldsymbol{\theta},k')}, \tag{8}$$

where $p(k|\boldsymbol{\theta})$ is the prior probability $p(k) = c_k$. Further, assuming the spectral bins to be statistically independent, $p_y(\mathbf{y}|\boldsymbol{\theta},k)$ can be factorized as

$$p_y(\mathbf{y}|\boldsymbol{\theta},k) = \prod_{d=1}^{D} p_{y_d}(y_d|\boldsymbol{\theta},k), \tag{9}$$

where the $p_{y_d}(y_d|\boldsymbol{\theta},k)$ can be represented as marginal distributions of the $p(y_d, x_d, n_d|\boldsymbol{\theta},k)$:

$$p_{y_d}(y_d|\boldsymbol{\theta},k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(y_d, x_d, n_d|\boldsymbol{\theta},k)dx_d dn_d, \tag{10}$$

with

$$p(y_d, x_d, n_d|\boldsymbol{\theta},k) = p(y_d|x_d, n_d, \boldsymbol{\theta})p_x(x_d|k)p_n(n_d). \tag{11}$$

In equation (11) it was assumed that speech and noise are statistically independent. Now, using the occlusion model (1) we have

$$p(y_d|x_d, n_d, \boldsymbol{\theta}) = \begin{cases} \delta_{x_d}(y_d), & \theta_d = 1 \\ \delta_{n_d}(y_d), & \theta_d = 0 \end{cases}$$

for hard decision masks. In the soft decision case, where θ_d is the probability of x_d being observed, this becomes

$$p(y_d|x_d, n_d, \boldsymbol{\theta}) = \theta_d \delta_{x_d}(y_d) + (1-\theta_d)\delta_{n_d}(y_d). \tag{12}$$

Combining equations (10), (11) and (12) and bounding both x_d and n_d below by 0 and above by y_d we obtain

$$\begin{aligned} p_{y_d}(y_d|\boldsymbol{\theta},k) \quad = \quad & \theta_d p_x(y_d)\int_0^{y_d} p_n(n_d)dn_d \quad + \\ & (1-\theta_d)p_n(y_d)\int_0^{y_d} p_x(x_d)dx_d. \end{aligned} \tag{13}$$

If we further assume that n_d is uniformly distributed on $[0, y_d]$ we obtain Raj's result [8]:

$$p_{y_d}(y_d|\boldsymbol{\theta},k) = \theta_d p_x(y_d) + \frac{(1-\theta_d)}{y_d}\mathcal{P}(0 \leq x_d \leq y_d|k), \tag{14}$$

where $\mathcal{P}(0 \leq x_d \leq y_d|k) = \mathcal{C}(y_d; \mu_{k,d}, \sigma_{k,d}^2) - \mathcal{C}(0; \mu_{k,d}, \sigma_{k,d}^2)$. Note that the more general form in equation (13) of this derivation explicitly allows the use of prior knowledge about the noise. This might be applied to approaches like [8] where the distribution of the noise is known.

3. PARTICLE FILTER BASED SOFT-MASK ESTIMATION

A variety of different methods has been proposed for mask estimation, including *computational auditory scene analysis* (CASA) [7], spectral subtraction [6], the difference between cube root signal and noise energy [10], Bayesian classifiers as well as the Max-VQ algorithm [8]. We use a particle filter. In order to explain how it can be used for mask estimation, we briefly sketch its operation (for details see [5, 12]): the particle filter for noise tracking approximates the probability distribution $p(\mathbf{n}_t)$ of the noise spectrum at time t as a set of N weighted noise hypotheses $\mathbf{n}_t^{(j)}$:

$$\hat{p}(\mathbf{n}_t) = \sum_{j=1}^{N} \omega_t^{(j)} \delta_{\mathbf{n}_t^{(j)}}(\mathbf{n}_t),$$

where the $\omega_t^{(j)}$ are the weights. This distribution is updated at each time t, using the following procedure:

1. **Sampling:** The noise hypotheses are propagated forward in time according to a dynamical system model that is either *autoregressive* as in [5, 12], *dynamic autoregressive* as described in [13], or a transition based on Polyac averaging and feedback [13].
2. **Weight Evaluation:** The weights $\omega_t^{(j)}$ of the noise hypotheses $\mathbf{n}_t^{(j)}$ are evaluated according to

   $$p_y(\mathbf{y}_t|\mathbf{n}_t) = p_x(f(\mathbf{y}_t, \mathbf{n}_t))\left|\text{det}\left(\frac{df(\mathbf{y}_t, \mathbf{n}_t)}{d\mathbf{y}_t}\right)\right| \tag{15}$$

   where p_x is an auxiliary clean speech Gaussian mixture model and where $\hat{\mathbf{x}}_t = f(\mathbf{y}_t, \mathbf{n}_t)$ is a noise compensation function, typically $f(\mathbf{y}_t, \mathbf{n}_t) = \log(e^{\mathbf{y}_t} - e^{\mathbf{n}_t})$. The multiplication by the absolute Jacobian determinant in equation (15) is due to transforming the probability density from p_y to p_x. After, the weights are normalized through division by $\sum_{j=1}^{N}\omega_t^{(j)}$.
3. **Resampling:** The noise hypotheses are pruned by multiplying hypotheses that have a relatively high weight and removing hypotheses that have a relatively low weight.

Now the soft-mask can be estimated as follows: the noise compensation function $f(\mathbf{y}_t, \mathbf{n}_t)$ from the particle filter can be used to obtain an estimate of the clean speech distribution from $\hat{p}(\mathbf{n}_t)$:

$$\hat{p}(\mathbf{x}_t) = \sum_{j=1}^{N} \omega_t^{(j)} \delta_{\mathbf{x}_t^{(j)}}(\mathbf{x}_t)$$

with $\mathbf{x}_t^{(j)} = f(\mathbf{y}_t, \mathbf{n}_t^{(j)})$. Then, using $\hat{p}(\mathbf{n}_t)$ and $\hat{p}(\mathbf{x}_t)$ the probability

$$\mathcal{P}(x_{t,d} > n_{t,d}) = \int_{-\infty}^{\infty}\int_{n_{t,d}}^{\infty} p(x_{t,d})dx_{t,d}p(n_{t,d})dn_{t,d}$$

can be approximated by Monte Carlo integration:

$$\mathcal{P}(x_{t,d} > n_{t,d}) \approx \sum_{j=1}^{N}\sum_{j'=1}^{N} \omega_t^{(j)}\omega_t^{(j')}u(x_{t,d}^{(j)} - n_{t,d}^{(j)}), \tag{16}$$

where u is the unit step function. Alternatively, a sigmoid function can be used as in [10]. Mask smoothing—averaging neighboring bins of the mask—increased performance slightly.

4. EXPERIMENTS

In the experiments reported below we used the phase-averaged particle filter described in [14] with 100 particles. For the proposed algorithm, the soft-mask was estimated according to equation (16). Subsequently, the noisy speech features were first enhanced, using the statistical inference approach (SFA) described in [12], then reconstructed according to equations (7) and (6). The posterior probabilities $p(k|\mathbf{y}, \boldsymbol{\theta})$ were calculated according to equations (14), (9) and (8).

The feature extraction of our ASR system was based on *Mel frequency cepstral coefficients* (MFCC)s. After *cepstral mean subtraction* (CMS) with variance normalization, 15 consecutive MFCC features were concatenated and subsequently reduced by *linear discriminant analysis* (LDA) to obtain the final 42-dimensional feature. The decoder used in the experiments is based on the fast on-the-fly composition of weighted finite-state transducers (WFSTs), as described in [13, §8]. The triphone acoustic model was trained with 30 hours WSJ0 and 12 hours WSJCAM0 data, resulting in 1,743 fully continuous codebooks with a total of 70,308 Gaussians. The auxiliary 128 component clean speech gaussian mixture model, used by the particle filter as well as missing feature reconstruction, was trained on the same data set.

The proposed particle filter for combined speech feature enhancement and missing feature reconstruction (PFR) was evaluated through a series of automatic speech recognition experiments. These experiments were conducted on the close talking channel of speakers 16-25 of the *multi-channel Wall Street Journal audio visual* (MC-WSJ-AV) corpus [15]. The corresponding 352 utterances were artificially contaminated by adding noise from the NOISEX-92 [16] database at different *signal-to-noise ratios* (SNR)s. Table 1 shows the results in comparison to the baseline (no PF) as well as to the particle filter without reconstruction (PFE), for a first, unadapted speech recognition pass as well as an adapted pass using constrained *maximum likelihood linear regression* (MLLR) [17]. On the unadapted pass the PFR clearly outperformed both the PFE and the baseline. The greatest gain was achieved on factory noise where, at 15 dB, the WER of the PFR was 26.1% lower than the baseline, compared to 17.0% for the PFE. On the adapted pass the results were not as clear: though the PFR always performed better than the baseline, it sometimes performed worse than the PFE. For 15dB destroyer engine and factory noise the PFR showed an additional gain of 3% and 4% relative over the PFE. For 10dB factory noise it performed worst — 2% (relative) worse than the PFE.

		5 dB		10 dB		15 dB	
noise	PF	1st	2nd.	1st	2nd	1st	2nd
destroyer	none	91.7	79.1	81.0	56.0	70.3	36.9
	PFE	86.7	72.0	73.2	47.9	63.3	33.1
	PFR	84.5	71.5	69.9	47.5	58.8	32.1
leopard	none	58.7	29.0	49.5	22.4	45.4	20.1
	PFE	54.8	26.0	47.1	20.6	41.1	20.3
	PFR	49.0	26.3	42.2	20.0	40.5	19.8
factory2	none	75.7	53.4	63.7	34.5	55.2	25.8
	PFE	70.6	49.9	53.6	31.1	45.8	24.2
	PFR	66.6	50.6	51.8	32.3	40.9	23.2

Table 1. *Word error rates* (WER)s for the particle filter (PFE), the particle filter with missing feature reconstrution (PFR) and the baseline (none) on the unadapted (1st) and adapted (2nd) pass. For clean speech the WER was 41.9% and 20.5% respectively.

5. CONCLUSIONS

We have motivated why MMSE noise compensation and missing feature reconstruction should be combined and shown how the particle filter can be used for soft-mask estimation. The good results on the unadapted pass lead us to believe that this approach is worth to be further pursued.

6. ACKNOWLEDGMENTS

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG), under the research training network IRTG 715 ”Language Technology and Cognitive Systems”.

7. REFERENCES

[1] P.J. Moreno, B. Raj, and R.M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” *Proc. ICASSP*, May 1996.

[2] N. S. Kim, “Nonstationary environment compensation based on sequential estimation,” *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 57–59, Mar. 1998.

[3] N. S. Kim, “IMM-based estimation for slowly evolving environments,” *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 146–149, June 1998.

[4] K. Yao and S. Nakamura, “Sequential noise compensation by sequential Monte Carlo method,” *Advances in Neural Information Processing Systems*, vol. 14, 2002.

[5] B. Raj, R. Singh, and R. Stern, “On tracking noise with linear dynamical system models,” *Proc. ICASSP*, May 2004.

[6] A. Drygajlo and M. El-Maliki, “Speaker verification in noisy environments with combined spectral subtraction and missing feature theory,” *Proc. ICASSP*, May 1998.

[7] M. Cooke, A. Morris, and P. Green, “Missing data techniques for robust speech recognition,” *Proc. ICASSP*, Apr. 1997.

[8] B. Raj and R. Singh, “Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition,” *Proc. ASRU*, Nov. 2005.

[9] P. Renevey, *Speech Recognition in Noisy Conditions Using Missing Feature Approach*, EPFL Lausanne, Lausanne, Switzerland, 2000.

[10] J. Barker, L. Josifovski, M. Cooke, and P. Green, “Soft decisions in missing data techniques for robust automatic speech recognition,” *Proc. ICSLP*, Oct. 2000.

[11] J. K. Patel and C. B. Read, *Handbook of the Normal Distribution*, Marcel Dekker Inc., New York, U.S.A., 1996.

[12] F. Faubel and M. Wölfel, “Overcoming the vector Taylor series approximation in speech feature enhancement - a particle filter approach,” *Proc. ICASSP*, Apr. 2007.

[13] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, New York, 2008.

[14] F. Faubel, J. McDonough, and D. Klakow, “A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain,” *Proc. Interspeech*, Sept. 2008.

[15] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, “The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments,” *Proc. ASRU*, Nov. 2005.

[16] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, pp. 247–251, 1993.

[17] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.