ROBUST SPEECH RECOGNITION USING CEPSTRAL DOMAIN MISSING DATA TECHNIQUES AND NOISY MASKS

Hugo Van hamme

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
hugo.vanhamme@esat.kuleuven.ac.be

ABSTRACT

Missing Data Techniques (MDT) have shown to be an effective method for curing the performance degradation of HMM-based speech recognition systems operating on noisy signals. However, a major drawback of the approach is that MDT requires that the acoustic model be expressed as a mixture of diagonal Gaussians in the log-spectral domain, whereas a higher accuracy can be obtained with Gaussian mixtures in the cepstral domain. This paper describes a recognizer based on the recently described cepstral-domain MDT approach using missing data masks computed from the noisy signal. It exploits a novel decision criterion that integrates harmonicity with signal-to-noise ratio and which makes minimal assumptions on the noise. The system is shown to exhibit a recognition accuracy that is comparable to the ETSI Advanced Front-End reference.

1. INTRODUCTION

Additive noise leads to a deterioration of speech recognition accuracy due to a mismatch between the noisy feature vector statistics and the speech models. Missing Data Techniques (MDT) have shown to be effective in reducing this mismatch for both bandlimited and wideband noise. In the MDT approach, features are labeled as either missing or reliable. The latter are considered to be free of noise and are used as such in the evaluation of an acoustic model trained on clean speech. Missing features on the other hand are either removed (marginalisation) or their value is inferred from the reliable features using the HMM state distribution as a prior (data imputation), see [1]. In this paper, the data imputation method will be applied.

A major drawback of the missing data approach is that it generally requires that the acoustic models are expressed in the (log-)spectral domain. However, speech recognition systems achieve a higher accuracy when the acoustic model is expressed in the cepstral domain and if velocity and acceleration features are used. Alternatively, a linear transform of the log-spectra, such as an LDA, can replace the cepstral representation. Recently, it was shown that Missing Data Techniques can be applied in the cepstral domain or linear transform domain as well [2], which solves the accuracy and robustness loss associated with the sub-optimal HMM emission density representation in the spectral domain.

Whereas [2] described the approach to acoustic modeling using MDT in the cepstral domain, the Missing Data Detector (MDD) was still idealized by using "oracle" or *a priori* masks derived from knowledge of the clean speech and the noise. In this paper, the noise masks are derived from the noisy signal based on weak assumptions about the noise. This paper proves that MDT systems can be competitive to carefully designed front-ends incorporating noise suppression techniques. The MDD presented in this paper assumes that (voiced) speech is the dominant harmonic signal component and hence the signal is decomposed into harmonically related plus random components. Though this "harmonicity" has been exploited before to build MDD's [3], the present approach integrates harmonicity and signal-to-noise ratio (SNR) through signal processing.

This paper is organized as follows. Section 2 briefly restates the missing data approach in the cepstral domain. The algorithm to decompose the signal into a harmonic and a random component is explained in section 3, while section 4 describes how this decomposition is applied to build missing data masks. Section 5 describes a method for additional noise reduction. The experiments on the AURORA-2 database are presented in section 6. Finally, section 7 concludes and describes how the present work will be carried forward.

2. MISSING DATA TECHNIQUES IN THE CEPSTRAL DOMAIN

In this paper, we focus on robustness to additive noise, while unknown filtering will not be considered. Although MDT can be applied to static and dynamic cepstra jointly [2], for the sake of computational simplicity, MDT is applied to static cepstra only. It was shown in [2] that MDT compensation of the velocity and acceleration features leads to only a moderate accuracy improvement.

The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture acoustic models. In the front-end, a low-resolution MEL spectral representation is computed by a filter bank through windowing, framing, FFT and filter bank integration. At frame t, the output of the filter bank with center frequency f will be denoted by $|Y_t(f)|$, $|S_t(f)|$ and $|N_t(f)|$ for the noisy speech, clean speech and noise respectively. The log-MEL-spectral noisy features $\mathbf{y}_t$ are then obtained by stacking $\log(|Y_t(f)|)$ for all filter banks in a vector, and likewise for $\mathbf{s}_t$ and $\mathbf{n}_t$. In missing data theory, it is now observed that

$$\mathbf{y}_t \approx \max(\mathbf{s}_t, \mathbf{n}_t) \qquad (1)$$

In [2] it was shown that the evaluation of the i-th Gaussian mixture component in the HMM emission density model of state q should be replaced by the non-negative least squares problem in the variable $\mathbf{x}$:

$$\begin{bmatrix} \boldsymbol{\Sigma}_{iq}^{-\frac{1}{2}}\mathbf{C}_u \\ \lambda\,\boldsymbol{\Sigma}_{a,iq}^{-\frac{1}{2}} \end{bmatrix}\mathbf{x} = \begin{bmatrix} \boldsymbol{\Sigma}_{iq}^{-\frac{1}{2}}\left(\mathbf{C}\mathbf{y}_t - \boldsymbol{\mu}_{iq}\right) \\ \lambda\,\boldsymbol{\Sigma}_{a,iq}^{-\frac{1}{2}}\left(\mathbf{y}_t - \boldsymbol{\mu}_{a,iq}\right) \end{bmatrix} \text{ with } \mathbf{x} \geq \mathbf{0}$$

where $\mathbf{C}$ is the (truncated) DCT transform matrix, $\boldsymbol{\mu}_{iq}$ and $\boldsymbol{\Sigma}_{iq}$ are the state cepstral mean and diagonal covariance of $\mathbf{C}\mathbf{s}_t$ and $\boldsymbol{\mu}_{a,iq}$ and $\boldsymbol{\Sigma}_{a,iq}$, are the mean and diagonal covariance of $\mathbf{s}_t$. The subscript u denotes those matrix columns corresponding to the unreliable components of $\mathbf{y}_t$, and λ is a non-critical regularization constant. The solution $\mathbf{x}$ is subtracted from the unreliable components of $\mathbf{y}_t$ to find the maximum likelihood clean speech estimate, which is optimal for Gaussian i of state q.

3. HARMONIC DECOMPOSITION

In order to determine which signal components are caused by noise and which are due to speech, the property that speech contains voiced segments composed of harmonically related components with "slowly" varying pitch is used. The underlying idea is that the harmonic part will be dominated by the speech rather than the noise. During voiceless speech, this approach may lead to poor decisions.

To decompose the signal into its harmonically related components, a pitch estimate is first computed. To this end a subharmonic summation method inspired by [4] is augmented with a dynamic programming algorithm to suppress doubling and halving errors. The signal is subsequently framed in overlapping segments with a length of two pitch periods and a single period of frame shift. Let t denote the pitch period index, then the noisy speech signal is written as:

$$y_t(n) = h_t(n) + r_t(n) \quad \text{with} \quad 0 \leq n < N_t$$

where N_t is the estimate of the double pitch period, as given by the closest frame of the subharmonic pitch estimator and $r_t(n)$ is the random signal component. The harmonic part is modeled as:

$$h_t(n) = \left(1 + \frac{c_t n}{N_t}\right)\left[\sum_{k=0}^{K_t} a_{k,t}\cos\left(2\pi f_{0,t} k n\right) + \sum_{k=1}^{K_t} b_{k,t}\,sin\left(2\pi f_{0,t} k n\right)\right] \quad (2)$$

where $f_{0,t}$ is the pitch estimate for segment t and the number of harmonics K_t is the largest integer such that $f_{0,t}K_t$<0.5. The inclusion of the linear modulation function $(1+c_t n/N_t)$ accounts for the change in speech amplitude that is observed over a length of N_t samples. This modulation parameter together with the pitch frequency and the harmonic amplitude parameters $a_{k,t}$ and $b_{k,t}$ are estimated as follows. Define the matrices $\mathbf{C}$ and $\mathbf{S}$ with entries $C_{nk} = \cos\left(2\pi f_{0,t} kn\right)$ and $S_{nk} = sin\left(2\pi f_{0,t} kn\right)$ with $1 \leq k \leq K_t$ and $0 \leq n < N_t$. Furthermore, let $'$ denote matrix transpose, $\mathbf{h}_t = [h_t(0), \ldots, h_t(N_t\text{-}1)]'$, $\mathbf{a}_t = [a_{0,t}, a_{1,t}, \ldots, b_{1,t}, \ldots]'$, and let $\mathbf{e}_t$ be a column vector of N_t ones. Finally, with $\mathbf{A}_t$ the diagonal matrix $diag\left(1, 1+c_t/N_t, 1+2c_t/N_t, \ldots, 1+(N_t-1)c_t/N_t\right)$ and $\mathbf{E}_t = \mathbf{A}_t\left[\mathbf{e}_t \quad \mathbf{C}_t \quad \mathbf{S}_t\right]$, the harmonic signal component is expressed as $\mathbf{h}_t = \mathbf{E}_t\mathbf{a}_t$. The parameters $\mathbf{a}_t$, $f_{0,t}$ and c_t will now be estimated in the least squares sense by minimizing $(\mathbf{y}_t\text{-}\mathbf{E}_t\mathbf{a}_t)'(\mathbf{y}_t\text{-}\mathbf{E}_t\mathbf{a}_t)$. For each choice of the scalar parameters $f_{0,t}$ and c_t, the matrix $\mathbf{E}_t$ is fixed and the estimation of $\mathbf{a}_t$ is a linear least squares problem with solution $\hat{\mathbf{a}}_t = \mathbf{R}_t^{-1}\mathbf{Q}_t'\mathbf{y}_t$ where $\mathbf{Q}_t\mathbf{R}_t = \mathbf{E}_t$ is the QR decomposition of $\mathbf{E}_t$ (with $\mathbf{R}_t$ square and $\mathbf{Q}_t$ is homomorphous to $\mathbf{E}_t$).

Substitution of $\hat{\mathbf{a}}_t$ in the error function yields the simplified cost $L(f_{0,t}, c_t) = \mathbf{d}'(f_{0,t}, c_t)\mathbf{d}(f_{0,t}, c_t)$ with $\mathbf{d}(f_{0,t}, c_t) = \mathbf{P}_t^{\perp}\mathbf{y}_t$ where $\mathbf{P}_t^{\perp} = \mathbf{I} - \mathbf{Q}_t\mathbf{Q}_t'$ is the projection matrix onto the null-space of $\mathbf{E}_t$. The cost can be minimized iteratively with updates of the parameter vector $[f_{0,t}\ c_t]'$ equal to $-\left(\nabla^2 L\right)^{-1}\nabla L$ where the gradient is found as $\nabla L = 2\mathbf{d}'\nabla\mathbf{d}$ and the Hessian is computed using the Levenberg-Marquardt approximation:

$$\nabla^2 L \approx 2\,diag(1+\varepsilon, 1+\varepsilon)\,\nabla\mathbf{d}'\nabla\mathbf{d} \text{ with } 0 \leq \varepsilon \ll 1.$$

With some algebra, the partial derivatives in $\nabla\mathbf{d}$ are found to be

$$\frac{\partial\mathbf{d}}{\partial f_{0,t}} = -\mathbf{Q}_t\left(\mathbf{R}_t^{-1}\right)'\frac{\partial\mathbf{E}_t'}{\partial f_{0,t}}\mathbf{P}_t^{\perp}\mathbf{y}_t - \mathbf{P}_t^{\perp}\frac{\partial\mathbf{E}_t}{\partial f_{0,t}}\hat{\mathbf{a}}_t$$

and likewise for the partial derivative with respect to c_t. The initial pitch value is set to the subharmonic summation pitch estimate and the initial value for c_t is zero. In our experiments, no examples of divergence have been observed while 2 to 3 iterations suffice. Because the harmonic summation method estimates the pitch from a low-pass spectrum, the recursive updates are required to find a good spectral fit for the higher harmonics. Together with the modulation model, this constitutes a refinement over the decomposition method presented in [5].

For each segment t, the estimated parameters are now plugged into (2), which is evaluated over the central pitch period of the segment. Subsequently, the harmonic signal part is formed by concatenation of the central segments. The random part is the difference between the orginal and the harmonic signal. The filter bank outputs computed on these harmonic and random components will be denoted by $|H_t(f)|$ and $|R_t(f)|$ respectively.

4. MISSING DATA DETECTOR

A crucial component of a speech recognizer based on missing data techniques is the missing data detector (MDD). The ideal MDD decision criterion labels the output of the filter bank at frequency f and frame t as missing if:

$$|S_t(f)|^2 < |N_t(f)|^2 \quad (3)$$

Since the phase relation between speech and noise is unknown, the expected value of both sides of (3) is taken. By using the statistical independence of speech and noise, one obtains:

$$2E\left\{|S_t(f)|^2\right\} < E\left\{|S_t(f)|^2\right\} + E\left\{|N_t(f)|^2\right\} = E\left\{|Y_t(f)|^2\right\}$$

Neglecting the correlation between H and R, this becomes:

$$2E\left\{|S_t(f)|^2\right\} < E\left\{|H_t(f)|^2\right\} + E\left\{|R_t(f)|^2\right\}$$

However, only one observation per frame is available, such that the expected values are estimated by their instantaneous values. The clean speech spectrum is estimated as

$$|S_t(f)|^2 = \gamma_t(f,q)|H_t(f)|^2 \quad (4)$$

for some time, frequency and state (or Gaussian)-dependent gain $\gamma_t(f,q)$. Hence the MDD criterion becomes:

$$\left(2\gamma_t(f,q) - 1\right)|H_t(f)|^2 < |R_t(f)|^2 \quad (5)$$

Although a gain estimate γ = 1 yields usable "harmonicity" masks, a better recognition accuracy was observed if γ is related to a signal-to-noise-derived estimate as follows (for brevity, the

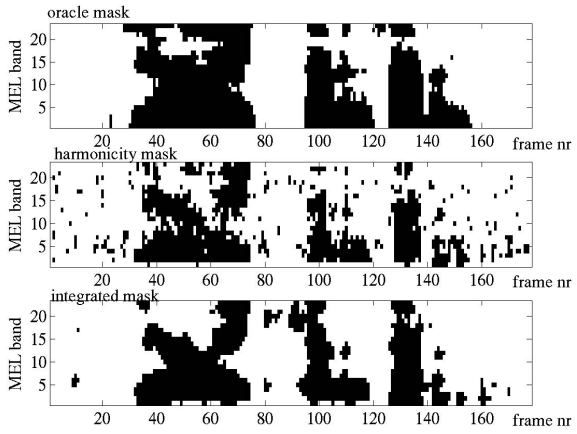

Figure 1: *Oracle, harmonicity and integrated mask for train noise at 10dB SNR.*

dependency on t, f and q is dropped in the sequel)

$$2\gamma - 1 = \frac{2|S|^2 - |H|^2}{|H|^2} \approx \frac{2\gamma^{(S)}|H^{(S)}|^2 - |H^{(S)}|^2 - |H^{(N)}|^2}{|H|^2}$$

$$\approx \left(2\gamma^{(S)} - 1\right) - 2\gamma^{(S)}|H^{(N)}|^2 \Big/ |H|^2$$

where superscript (S) and (N) denote the variable computed on clean speech or noise only. Although prior knowledge about the state can now be introduced in this estimator, the following state-independent approximations were selected:

- $\gamma^{(S)} = 1$, which is realistic for voiced speech
- $|H_t^{(N)}| \approx |R_t^{(N)}| \approx \alpha \min_{t-L \le s \le t+L} |R_s|$ with $\alpha > 0$

The approximation of $|H_t^{(N)}|$ by $|R_t^{(N)}|$ can easily be understood from the harmonic decomposition method. To this end, first assume that $f_{0,t}$ is fixed and that $c_t = 0$, i.e. only the amplitude parameters are estimated. Due to the choice of N_t as twice the period, the matrix $\mathbf{E}_t$ becomes orthogonal and $a_{t,k}$ ($b_{t,k}$) are found as the real (imaginary) part of the even-numbered lines of the IDFT spectrum computed on N_t data points. The odd-numbered spectral lines form the remainder or random signal part. Hence, for random (non-harmonic) noise, $|H|$ and $|R|$ are obtained by MEL-integration of interleaved subsampling of the noise spectrum. With $f_{0,t}$ and c_t estimated in the least squares sense, the energy in the harmonic part will increase slightly at the expense of the random component. Hence, $|R_t^{(N)}|$ will underestimate $|H_t^{(N)}|$. Finally, $|R_t^{(N)}|$ is estimated from the minimum of $|R|$ over a window of $2L+1$ frames (here $L = 10$).

From (5) it is clear that harmonicity and SNR evidence are integrated in a global decision. Like in [3], balancing MDD misclassification errors of the 1st and 2nd kind is achieved by multiplying the right-hand-side of equation (5) by a "margin" (here –10 dB). This threshold is raised to 0 dB for silence states as a rudimentary implementation of state-dependent γ-estimates. Figure 1 illustrates the superiority of the integrated masks over the harmonicity masks when their margins are chosen to yield equal ratios of errors of the 1st and 2nd kind.

5. NOISE REDUCTION

Especially when the speech only slightly dominates the noise, the max-approximation (1) involves a non-negligible error in the log-spectral domain. When speech and noise are equal in amplitude, the $|Y|$ will on average be 3 dB greater than $|S|$. An option would be to use (4) as an estimate of the clean speech. However, the approximations motivated above also imply that $\gamma \le 1$, so $|S|$ would be underestimated. Instead, the ratio of the left and right hand side of (5) (the "decision margin" of the MDD) is used to define a gain function $g_t(f)$ that is multiplied with $|Y|$. To avoid overcompensation, $g_t(f)$ is soft-limited between –6 dB and 0 dB. Hence, areas in the time-frequency plane that are prominently labeled as speech are unaffected, the noisy parts are suppressed by 6 dB and the desired 3 dB suppression are obtained when speech and noise have a comparable magnitude. A major difference with spectral subtraction, where a similar gain function is used, is that in this approach, we do not attempt to recover the clean speech in case it is masked by the noise.

This noise reduction step is also required to enhance the velocity and acceleration features, such that a fair comparison relative to the next section's baseline can be obtained, in which compensates the dynamic parameters as well.

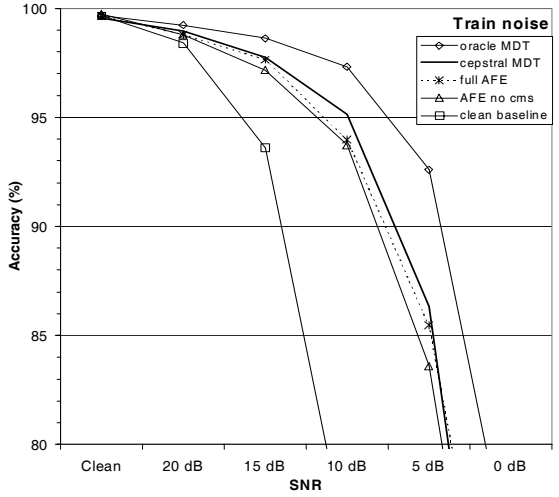

Figure 2: *Comparison chart for AURORA-2, Set A, Noise 1.*

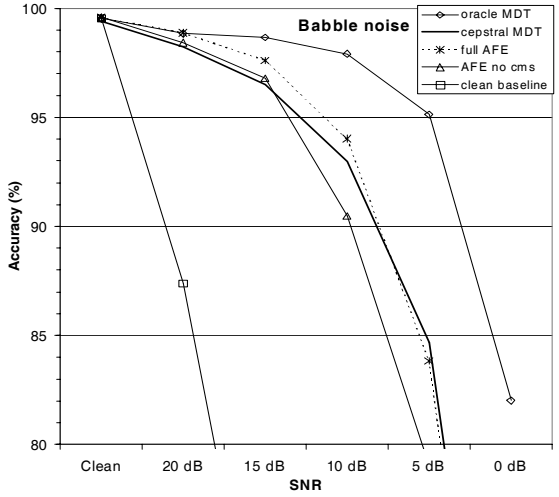

Figure 3: *Comparison chart for AURORA-2, Set A, Noise 2.*

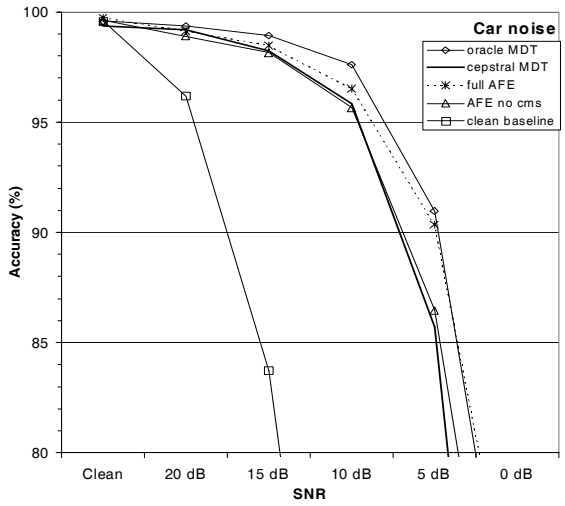

Figure 4: *comparison chart for AURORA-2, Set A, Noise 3.*

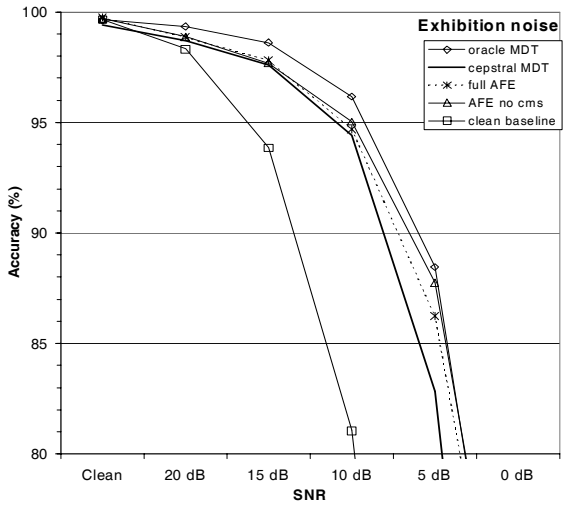

Figure 5: *comparison chart for AURORA-2, Set A, Noise 4.*

6. EXPERIMENTS

The above approach is evaluated on the AURORA-2 continuous digit recognition task using the complex back-end. The evaluation is limited to test set A, since channel mismatch (test set C) is beyond the scope of this paper and also knowledge about the noise is not exploited during the training process (test set B). This configuration consists of an HMM Gaussian mixture architecture with 16 states per digit and 20 Gaussians per state. The optional inter-word silence is modeled by 1 or 3 states with 36 Gaussians per state, while leading and trailing silence have 3 states. The total number of Gaussians is 3628. The front-end of the MDT system is the ETSI STQ WI-007 standard, a textbook MFCC feature extraction method without cepstral mean normalization. Velocity and acceleration features are computed using the HTK default regression formulae.

The accuracy results are presented in figure 2 through 5 for the four noise types of the test set. The curve labeled "full AFE" is obtained using a reference system composed of a back-end of the same complexity as above and the "Advanced Front-End" (AFE) as described by ETSI STQ WI-008 standard [6], where speech enhancement is performed by a two-stage Wiener filtering and subsequent waveform processing. Since the present focus does not include compensation for unknown filtering, the "Blind Equalization" was removed from the AFE baseline and the system was configured not to use the voice activation detection. Because there is no filtering mismatch between test and training, the accuracy impact of this modification is minor, as is witnessed by the curves labeled "AFE no cms". The curve labeled "oracle MDT" uses (3) as a decision criterion and shows the further potential of the cepstral MDT method. The results of the system described above are given as "cepstral MDT". These results compare favorably to the "AFE no cms" reference. The worst results are obtained for the exhibition noise, which contains whistling in many files. This noise violates the assumption that the harmonic signal is due to speech. Advanced auditory scene analysis or even simple pitch constraints could alliviate the problem. Finally, a reference system using the ETSI WI-007 front-end and clean speech model but without MDT compensation is shown as "clean baseline".

7. CONCLUSIONS

A speech recognizer based on Missing Data Techniques in the cepstral domain was described. Unlike in previous work, the missing data masks were computed from the noisy signal based on an original method involving harmonic decomposition. Harmonic components were assumed to originate from the speech only. No long-term noise averages were used. The experimental evidence shows that cepstral MDT systems can achieve a degree of robustness to additive noise that is comparable to the ETSI Advanced Front-End.

Further work will include: HMM-state-dependent estimates of γ, soft decisions or fuzzy masks in the cepstral MDT approach and compensation for unknown filtering.

8. REFERENCES

[1] M. Cook, Ph. Green, L. Josifovski, and A. Vizinho "Robust automatic speech recognition with missing and unreliable acoustic data". *Speech Communication 34 (2001)*, pp. 267-285.

[2] H. Van hamme, "Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain," *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 3089-3092.

[3] J. Barker, M. Cooke, and Ph. Green, "Robust ASR Based On Clean Speech Models: An Evaluation of Missing Data Techniques For Connected Digit Recognition In Noise", *in Proc. EUROSPEECH*, Denmark, September 2001, pp. 213-216.

[4] D. Hermes, "Measurement of Pitch by Subharmonic Summation", *J. Acoust. Soc. Am.* **83** (1), January 1988, pp. 257-264.

[5] M. Seltzer, J. Droppo, and A. Acero, "A Harmonic-Model-Based Front End For Robust Speech Recognition", *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 1277-1280.

[6] "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm", *ETSI ES 202 050 v1.1.1 (2002-10).*