Kalman and Unscented Kalman Filter Feature Enhancement for Noise Robust ASR

Veronique Stouten[‡]*, Hugo Van hamme, Patrick Wambacq*

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{vstouten,hvanhamm,wambacq}@esat.kuleuven.ac.be

10.21437/Interspeech.2005-227

Abstract

Model-based feature enhancement is an ASR front-end technique to increase the robustness of the recogniser in noisy environments. However, its MMSE-estimates of the clean speech feature vectors are based only on the static components at the current frame. In this paper, we show how the Kalman filter framework can be seen as a natural extension that incorporates both the current and the previous frames in the enhancement process. Because multiple Kalman filters are run in parallel, the global clean speech estimate is given by a weighted linear combination of the individual MMSE-estimates. Also, the unscented transformation is considered to avoid the linearisation of the cepstral domain observation equation. We present experimental results on the Aurora2 database for both the multi-modal Kalman and the unscented Kalman filter feature enhancement.

1. Introduction

The Kalman filter implements a predictor-corrector type estimator and can be used to track the state of a system, while minimising the trace of the estimated error covariance matrix. The approach is a well studied topic in the speech enhancement context [1, 2, 3]. However, the application of this algorithm in the context of automatic speech recognition (ASR) is limited. In [4] for instance, the Kalman filter is used to track the non-stationary noise characteristics.

The advantage of a Kalman filter is that not only the current, but also the previous observations are taken into account. Up to a linear transformation this can be made equivalent to extending the static features with the velocity and the acceleration features. In this context, the Kalman filter can be seen as a natural extension of the model-based feature enhancement [5]. Similar to the clean speech HMM used in the latter approach, prior knowledge about the clean speech is used by running multiple Kalman filters in parallel (multi-modal Kalman filter). Each filter has its own auto-regressive state transition matrix that relates the states at different frames. The global clean speech estimate is obtained by combining all the individual estimates according to the posterior probability of their model (filter).

One of the Kalman filter extensions, namely the unscented Kalman filter (UKF), was proposed by Julier and Uhlmann [6]. In this technique, the unscented transformation is used to avoid the Vector Taylor Series linearisation of the observation equation. Instead of approximating the non-linear function, it approximates the state probability distribution. By propagating a small set of deterministically chosen (sigma) points through the true non-linear system, the posterior mean and covariance of a Gaussian random variable can be captured accurately to the second order of the non-linearity.

In this paper, we show how both the Kalman and the unscented Kalman filter can be applied in the context of noisy speech feature enhancement for noise robust ASR. In section 2 we describe how the multi-modal Kalman filter is used to track the clean speech cepstral feature vector and illustrate its similarity to the MBFE-algorithm. Section 3 is devoted to the unscented Kalman filter. The performance of both algorithms is compared in section 4 in terms of recognition accuracy on the Aurora2 database. Conclusions can be found in section 5.

2. Kalman filter

2.1. Theory and notations

The Kalman filter addresses the general problem of trying to estimate the state $x \in \mathbb{R}^d$ of a discrete-time process that is governed by a set of linear state-space equations :

$$x_{t+1} = Ax_t + w_t \quad (1)$$

$$y_t = Bx_t + v_t \quad (2)$$

The $(d \times d)$ matrix A relates the state at the current time step t to the state at the next time step $(t+1)$ in the absence of process noise. The $(m \times d)$ matrix B relates the state to the observation $y \in \mathbb{R}^m$. The random variables w_t and v_t represent the process and measurement noise, respectively. They are assumed to be independent, white and with Gaussian probability distributions :

$$p[w_t] = N(0, W) \quad (3)$$

$$p[v_t] = N(0, V) \quad (4)$$

The Kalman filter algorithm then consists of the sequential application of two kinds of equations. The *time update* ('prediction') equations project the current state and error covariance estimates forward in time to obtain the a priori estimates for the next time step :

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1} \quad (5)$$

$$P_{t|t-1} = AP_{t-1|t-1}A' + W \quad (6)$$

in which prime denotes matrix transpose. The *measurement update* ('correction') equations incorporate a new observation into the a priori estimate to obtain an improved a posteriori estimate :

$$K_t = P_{t|t-1}B'\left(BP_{t|t-1}B' + V\right)^{-1} \quad (7)$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t\left(y_t - B\hat{x}_{t|t-1}\right) \quad (8)$$

$$P_{t|t} = (I - K_tB)\,P_{t|t-1} \quad (9)$$

Matrix K is called the Kalman gain.

[‡] Veronique Stouten is a Research Assistant of the Fund for Scientific Research - Flanders (Belgium) (F.W.O. - Vlaanderen).

2.2. Multi-modal filters

In this paper, the state vector x_t is the concatenation of the (unknown) current and $(M-1)$ previous clean speech cepstral feature vectors (static features only) :

$$x_t = \left[\begin{array}{cccc} s'_t & s'_{t-1} & \dots & s'_{t-M+1} \end{array} \right]' \quad (10)$$

Prior knowledge about the clean speech is incorporated by instantiating M^y independent Kalman filters that are run in parallel (multi-modal, alias multiple model). Each filter i has its own parameters A, B, W and V that are estimated offline on training data. To this end, the data are aligned to a clean speech GMM (that covers the clean speech acoustic space) and for each Gaussian $N(\mu_i^s, \Sigma_i^s)$, a set of Kalman filter parameters is calculated in least squares sense. We assume an Mth order auto-regressive (AR) relation between the clean speech cepstral vectors, such that the matrix A for each filter i has the form :

$$A_i = \left[\begin{array}{cccccc} \tilde{A}_1 & \tilde{A}_2 & \dots & \tilde{A}_{M-1} & \tilde{A}_M & \tilde{a}_0 \\ I & 0 & \dots & 0 & 0 & 0 \\ 0 & I & \dots & 0 & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & \dots & I & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{array} \right]_i \quad (11)$$

in which I is the identity matrix and $\tilde{a}_0$ allows for a constant offset such that w_t has mean zero. The diagonal submatrices $\tilde{A}$ and $\tilde{a}_0$ are trained in least square sense on clean speech training data. The process noise covariance matrix W is given by :

$$\begin{aligned} W_i &= cov(x_{t+1}|x_t) \\ &= \left[\begin{array}{ccccc} \tilde{W}' & 0 & \dots & 0 & 0 \end{array} \right]'_i \end{aligned} \quad (12)$$

with

$$\tilde{W}_i = cov(s_{t+1}, s_{t+1}) - \left[\tilde{A}_1 \dots \tilde{A}_M \right]_i cov(x_t, s_{t+1}) \quad (13)$$

In general, the observation equation (2) that relates the noisy speech cepstral vector y_t, the clean speech s_t, the additive noise n_t and the channel h is non-linear :

$$\begin{aligned} y_t &\approx f(s_t, n_t, h) \\ &\approx C \log\left(\exp\left(C^{-1}(s_t + h)\right) + \exp\left(C^{-1} n_t\right)\right) \end{aligned} \quad (14)$$

in which C^{-1} denotes the inverse of the DCT-matrix C. In the Kalman filter algorithm, it is approximated by a first order Vector Taylor Series around the corresponding means of clean speech μ_i^s and noise μ^n :

$$y_t \approx f\left(\mu_i^s, \mu^n, h\right) + F_i\left(s_t - \mu_i^s\right) + G_i\left(n_t - \mu^n\right) \quad (15)$$

The first derivatives of $f(s, n, h)$ to s and n, respectively, are denoted by F_i and G_i. Matrix B from equation (2) then has the form :

$$B_i = \left[\begin{array}{ccccc} F_i & 0 & \dots & 0 & f(\mu_i^s, \mu^n, h) - F_i \mu_i^s \end{array} \right] \quad (16)$$

and the measurement noise covariance matrix is given by :

$$V_i = G_i \Sigma^n G_i' + \varphi \quad (17)$$

where the residual $r = y - f(s, n, h)$, caused by the phase difference between speech and noise, is assumed zero mean Gaussian $N(0, \varphi)$ [7].

At each time instant, the final combined estimate of the state vector is computed as the weighted combination of the corresponding a posteriori estimates $\hat{x}_{t|t}$ for each filter i. The weights are given by the likelihood of the filter model, given the observation.

2.3. Remarks

Since the system matrices of (1) and (2) are constant over time, it can be shown [8] that $P_{t|t}$ converges to a fixed matrix P for each Kalman filter i. This property allows a computationally tractable implementation of the recursive algorithm.

Also, it is interesting to note that the Kalman filter formulae for the degenerate case of a 0th order AR-model reduce to the MMSE-estimator from the MBFE-algorithm of [5]. Indeed, in this case the state-space equations for filter i reduce to :

$$\begin{aligned} x_{t+1} &= \tilde{a}_{0,i} + w_t & (18) \\ y_t &= F_i x_t + f\left(\mu_i^s, \mu^n, h\right) - F_i \mu_i^s + v_t & (19) \end{aligned}$$

and the measurement update becomes :

$$\begin{aligned} K_{t,i} &= W_i B_i' \left(B_i W_i B_i' + V_i\right)^{-1} & (20) \\ \hat{x}_{t|t} &= \tilde{a}_{0,i} + K_{t,i}\left(y_t - F_i \tilde{a}_{0,i} - f\left(\mu_i^s, \mu^n, h\right) + F_i \mu_i^s\right) & (21) \end{aligned}$$

By using $(\tilde{a}_{0,i} = \mu_i^s)$ and recognising that the upper left submatrix of $(W_i B_i')$ is $(\Sigma_i^s F_i')$, the similarity to the MBFE-equation :

$$\hat{x}_t = \mu_i^s + \Sigma_i^s F_i' \left(\Sigma_i^y\right)^{-1} \left(y_t - \mu_i^y\right) \quad (22)$$

becomes evident.

3. Unscented Kalman filter

To avoid the linearisation of (14) in the observation equation, the UKF has been proposed [6, 9]. The unscented transformation operates on a random variable which is the concatenation of the state x_t and the noise variables w_t, v_t. Let d_z be the dimension of this concatenated vector, λ a scaling parameter, weights $\eta_0 = \lambda/(d_z + \lambda)$ and $\eta_k = 1/(2d_z + 2\lambda)$ for $k = 1 \dots 2d_z$. In our experiments, λ was set to 0. Again, M^y filters are run in parallel, but for convenience of notation, we will drop the index i in this part. The UKF algorithm initialisation is given by :

$$\begin{aligned} \hat{z}_{0|0} &= \left[\begin{array}{ccc} x'_{0|0} & 0' & (\mu^n)' \end{array} \right]' \\ Q_{0|0} &= \left[\begin{array}{ccc} P_{0|0} & 0 & 0 \\ 0 & \tilde{W} & 0 \\ 0 & 0 & \Sigma^n \end{array} \right] \end{aligned} \quad (23)$$

For frame $t = 1 \dots T$, the update formulae are as follows. Let $\sqrt{(.)}$ denote the matrix square root. First calculate :

$$\begin{aligned} \hat{Z}^{\pm}_{t-1|t-1} &= \hat{z}_{t-1|t-1} \cdot [1 \dots 1] \pm \sqrt{(d_z + \lambda)\, Q_{t-1|t-1}} \\ \chi_{t-1|t-1} &= \left[\begin{array}{ccc} \hat{z}_{t-1|t-1} & \hat{Z}^{+}_{t-1|t-1} & \hat{Z}^{-}_{t-1|t-1} \end{array} \right] \end{aligned} \quad (24)$$

The $(1 + 2d_z)$ columns of χ are called the sigma points and represent the pdf of $\hat{z}$. Further let $\chi' = [\ (\chi^x)' \quad (\chi^w)' \quad (\chi^v)'\ ]$.

- Time update :

$$\begin{aligned} \chi^x_{t|t-1} &= A\, \chi^x_{t-1|t-1} + \chi^w_{t-1|t-1} \\ \hat{x}_{t|t-1} &= \sum_{k=0}^{2d_z} \eta_k\, \chi^x_{k,t|t-1} \\ P_{t|t-1} &= \sum_{k=0}^{2d_z} \eta_k \left(\chi^x_{k,t|t-1} - \hat{x}_{t|t-1}\right)\left(\chi^x_{k,t|t-1} - \hat{x}_{t|t-1}\right)' \\ \psi_{t|t-1} &= f\left(\chi^x_{t|t-1}, \chi^v_{t|t-1}, h\right) \\ \hat{y}_{t|t-1} &= \sum_{k=0}^{2d_z} \eta_k\, \psi_{k,t|t-1} \end{aligned} \quad (25)$$

- Measurement update :

$$
\begin{aligned}
P_{\hat{y}_t\hat{y}_t} &= \sum_{k=0}^{2d_z} \eta_k \left(\psi_{k,t|t-1} - \hat{y}_{t|t-1}\right)\left(\psi_{k,t|t-1} - \hat{y}_{t|t-1}\right)' \\
P_{\hat{x}_t\hat{y}_t} &= \sum_{k=0}^{2d_z} \eta_k \left(\chi^x_{k,t|t-1} - \hat{x}_{t|t-1}\right)\left(\psi_{k,t|t-1} - \hat{y}_{t|t-1}\right)' \\
K_t &= P_{\hat{x}_t\hat{y}_t}\left(P_{\hat{y}_t\hat{y}_t}\right)^{-1} \\
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t\left(y_t - \hat{y}_{t|t-1}\right) \\
P_{t|t} &= P_{t|t-1} - K_t\, P'_{\hat{x}_t\hat{y}_t} \qquad (26)
\end{aligned}
$$

Hence, the non-linearly transformed mean $\hat{y}$ and covariance $P_{\hat{y}\hat{y}}$ are estimated from the 'cloud' of transformed sigma points ψ. Again, the final combined estimate of the state vector is computed as the weighted combination of the corresponding a posteriori estimates $\hat{x}_{t|t}$ for each filter i.

4. Experiments

Recognition experiments are conducted on the Aurora2 digit recognition task. Features are extracted by the MFCC front-end, complying to the ETSI ES 201 108 standard without compression. In our experiments, 128 independent (unscented) Kalman filters are run in parallel. A channel estimate is calculated online by a recursive EM-algorithm [10]. Front-end estimates are evaluated by the complex backend recognition system, with whole word digit models trained on the Aurora2 clean speech training database using the HTK scripts with default settings (see [11]). The digit models have 16 emitting states with 20 Gaussians per state, while the silence model has 3 states with 36 Gaussians per state. Also, a one-state short pause model, tied with the middle state of the silence model, is used.

4.1. Order of the auto-regressive model

To determine the best order of the AR-model in (11), experiments are conducted on a selected subset of Aurora2. Table 1 presents the recognition accuracy obtained with a Kalman filter using a 0th, 1st, 2nd and 4th order AR-model. The 1st order AR-model achieves the best results for all noise conditions that are considered, except for the SNR-level of 20 dB.

	A N1 SNR10	A N3 SNR0	B N1 SNR20	B N4 SNR5	C N2 SNR15	Avg.
0th	95.30	58.10	99.23	82.60	96.07	86.26
1st	96.04	62.06	99.11	84.54	96.58	87.66
2nd	95.95	61.77	99.14	84.11	96.49	87.49
4th	95.55	61.02	99.14	83.46	96.37	87.10

Table 1: *Recognition accuracy after enhancement with a Kalman filter using a 0th, 1st, 2nd and 4th order AR-model.*

Also, the mean squared prediction error is compared for the different AR-models. As can be seen in figure 1, the largest decrease in normalised prediction error is found between the 1st and the 2nd order AR-model. Although the error still decreases (hence a better representation of speech is obtained) for higher order AR-models, the recognition accuracy does not. Because a significantly higher recognition accuracy is only found for the 1st order AR-model, the results presented in section 4.2 are obtained with a 1st order AR-model. This means that only the current frame is taken into account to predict the next state.

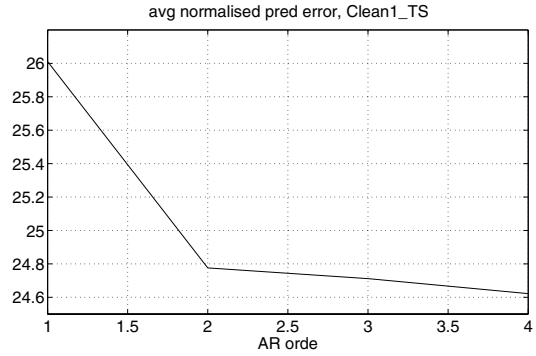

Figure 1: *Average of normalised MSE prediction error of AR-model versus model order. (Aurora2, clean1 test data).*

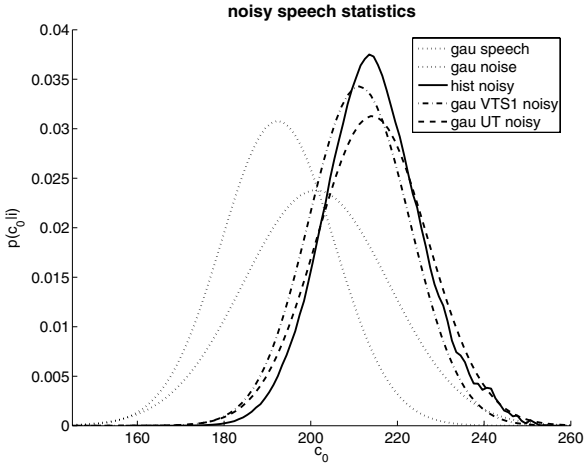

Figure 2: *Gaussian of clean speech and noise, corresponding histogram of noisy speech samples, Gaussian of noisy speech with 1st order VTS and with unscented transformation.*

4.2. Results

The recognition results are compared for the enhanced feature stream, obtained by the 1st order Kalman and unscented Kalman filter, respectively. The evaluation is done on the 4 noise types of test set A from the Aurora2 database. As a reference, table 2 presents the accuracy obtained with a 0th order Kalman filter, which is equivalent to the MBFE-algorithm of [10], and table 5 presents the results from the ETSI Advanced Front-End (AFE) standard [12].

As can be seen from table 3, it is beneficial to incorporate information from previous frames in the feature enhancement process. By using a 1st order Kalman filter, we can increase the average recognition accuracy from 87.20% to 88.55%. This result is comparable to the 88.56% that is obtained with the ETSI AFE standard (table 5), although in the latter technique less prior knowledge is used. Only for very low SNR-levels (0 dB) the AFE performs better than the Kalman filter, but for higher SNR-levels (between 5 dB and 20 dB) the accuracy of the proposed Kalman filter approach is superior.

Finally, table 4 presents the recognition results for the 1st order unscented Kalman filter. An average recognition accuracy of 88.04% is obtained, which is again higher than for the 0th order Kalman filter. Surprisingly, the unscented Kalman filter

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.20	99.12	99.22	99.01	99.14
15 dB	97.91	98.10	98.66	97.41	98.02
10 dB	95.30	94.38	95.79	93.61	94.77
5 dB	88.67	81.38	84.73	84.60	84.84
0 dB	68.62	48.67	58.10	61.49	59.22
Avg.	89.94	84.33	87.30	87.22	87.20

Table 2: *Recognition accuracy with a 0th order Kalman filter.*

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.23	99.12	99.40	99.11	99.22
15 dB	98.16	98.10	98.90	97.84	98.25
10 dB	96.04	94.95	96.54	94.48	95.50
5 dB	90.54	83.34	86.79	86.08	86.69
0 dB	71.32	53.05	62.06	65.87	63.08
Avg.	91.06	85.71	88.74	88.68	88.55

Table 3: *Recognition accuracy with a 1st order Kalman filter.*

is not able to improve the performance of the 1st order Kalman filter. Figure 2 shows the noisy speech statistics obtained with a Monte Carlo simulation, with a VTS linearisation and with the unscented transformation, respectively. Although the UT mean is more accurate, its variance seems overestimated, which might explain the lower performance.

5. Conclusions

In this paper, we have shown how the well-known Kalman filter can be applied to track the clean speech cepstral feature vector in the context of feature enhancement for noise robust ASR. The advantage of a Kalman filter is that not only the current, but also the previous frames are taken into account. To avoid the linearisation of the cepstral domain observation equation, the unscented transformation was considered.

Prior knowledge about the clean speech is used by running multiple (unscented) Kalman filters in parallel, each with their own auto-regressive state transition matrix. The global clean speech estimate is obtained by combining all the individual estimates according to the posterior probabilities. Interestingly, the Kalman filter formulae for a 0th order AR-model reduce to the MMSE-estimation that is used in model-based feature enhancement (MBFE).

Recognition results showed that the Kalman filter approach with a 1st order AR-model outperforms the MBFE-algorithm. Also, the accuracy of the proposed Kalman filter is superior to the ETSI AFE for SNR-levels between 5 dB and 20 dB. Although the mean squared prediction error of the AR-model still decreased for higher orders, increasing the order was not beneficial to the recognition accuracy. Also, the 1st order unscented Kalman filter performs better than the MBFE-algorithm, but not as good as the 1st order Kalman filter.

Future work includes the joint estimation of the speech and noise statistics. Instead of using only an AR-model for speech, the Kalman state vector can be a concatenation of speech and noise vectors.

6. Acknowledgement

This work was partly supported by 'Research Fund (Onderzoeksfonds) K.U.Leuven', project no. OT/03/32/TBA.

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	99.26	99.24	99.34	98.95	99.20
15 dB	98.37	98.31	98.69	97.62	98.25
10 dB	96.13	94.83	96.30	93.74	95.25
5 dB	89.50	82.38	85.92	85.34	85.78
0 dB	69.97	52.09	60.39	64.39	61.71
Avg.	90.65	85.37	88.13	88.01	88.04

Table 4: *Recognition accuracy with a 1st order UKF.*

	Subway	Babble	Car	Exhibit.	Avg.
20 dB	98.80	98.88	99.14	98.86	98.92
15 dB	97.64	97.61	98.48	97.81	97.89
10 dB	93.98	94.01	96.51	94.69	94.80
5 dB	85.47	83.83	90.36	86.24	86.48
0 dB	65.65	56.48	71.18	65.54	64.71
Avg.	88.31	86.16	91.13	88.63	88.56

Table 5: *Recognition accuracy with the ETSI AFE standard.*

7. References

[1] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Proc. ICASSP*, Dallas, Texas, Apr. 1987, vol. I, pp. 177–180.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE TSAP*, vol. 6, no. 4, pp. 373–385, 1998.

[3] V. Grancharov, J. Samuelsson, and W. Bastiaan Kleijn, "Improved kalman filtering for speech enhancement," in *Proc. ICASSP*, Philadelphia, Mar. 2005, pp. 1109–1112.

[4] N.S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Comm.*, vol. 37, pp. 231–248, 2002.

[5] V. Stouten, H. Van hamme, K. Demuynck, and P. Wambacq, "Robust speech recognition using model-based feature enhancement," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 17–20.

[6] S.J. Julier and J.K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. of the IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[7] V. Stouten, H. Van hamme, and P. Wambacq, "Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement," in *Proc. ICASSP*, Philadelphia, Mar. 2005, vol. I, pp. 433–436.

[8] B.M. Klein, *State Space Models for Exponential Family Data*, Ph.D. thesis, Univ. Southern Denmark, May 2003.

[9] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan, "The unscented particle filter," Technical Report CUED/F-INFENG/TR 380, Cambridge Univ. Engineering Department, Aug. 2000.

[10] V. Stouten, H. Van hamme, and P. Wambacq, "Joint removal of additive and convolutional noise with model-based feature enhancement," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 949–952.

[11] "http://htk.eng.cam.ac.uk/," .

[12] ETSI standard doc., "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)*.