Vector-Quantization based Mask Estimation for Missing Data Automatic Speech Recognition

Maarten Van Segbroeck[*], *Hugo Van hamme*

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{maarten.vansegbroeck,hugo.vanhamme}@esat.kuleuven.be

10.21437/Interspeech.2007-330

Abstract

The application of Missing Data Theory (MDT) has shown to improve the robustness of automatic speech recognition (ASR) systems. A crucial part in a MDT-based recognizer is the computation of the reliability masks from noisy data. To estimate accurate masks in environments with unknown, non-stationary noise statistics only weak assumptions can be made about the noise and we need to rely on a strong model for the speech. In this paper, we present a missing data detector that uses harmonicity in the noisy input signal and a vector quantizer (VQ) to confine speech models to a subspace. The resulting system can deal with additive and convolutional noise and shows promising results on the Aurora4 large vocabulary database.

Index Terms: speech recognition, noise robustness, missing data mask estimation, speech separation

1. Introduction

Additive noise results in a decrease in performance of ASR-systems due to the mismatch between the speech models (obtained in clean conditions) and the statistics of speech in the noisy test conditions. Missing Data Techniques have shown their effectiveness in reducing this mismatch. In the MDT approach, a missing data detector decides for each frame which features of the noisy speech are unreliable (missing) or reliable. The latter are treated as clean speech data in the acoustic models of the recognizer's backend. The missing features on the other hand are either marginalized or their value is estimated from the reliable data using the state distribution of the ASR as a prior (data imputation), see [1]. For reasons of accuracy, the ASR-system operates in a domain that is a linear transformation of log-spectra. Therefore, the data imputation technique needs to be extended to cover such linear transformations. An alternative MDT formulation was presented in [2] through the introduction of the PROSPECT features and will be briefly restated in section 2.

One of the main advantages of missing data techniques over other noise reduction methods is that less assumptions need to be made about the noise type. However, the estimation of the mask is often based on important assumptions about stationarity, signal-to-noise (SNR) ratio or statistical distribution of the noise. Hence, to separate speech from noise, more constraints should be placed on the speech model. Therefore, we must exploit the a-priori knowledge of the human voice, such as spectral characteristics, harmonicity, energy, voicing and onset. As shown in [3], a constrained subspace for the spectral shape of speech signals can be captured in a vector quantization codebook trained on features extracted from clean speech. If speech is corrupted by noise, simple nearest-neighbor decoding fails due to the mismatch between the codebook training set and the noisy test set. Previous approaches for a more noise robust decoding were reported in [4], using a more perceptually motivated distance measure and [5], including phase derivatives.

Voiced speech is characterized by its strong harmonicity arising from the presence of the pitch and its harmonics such that it can be decomposed into harmonic components at integer pitch multiples and the remaining aperiodic or random components. In the proposed VQ-system, we exploit this characteristic by training the codebook on the spectral features extracted from the harmonic and random part of the clean speech signal. During speech events, the decoding seeks to recover the original speech vector from the stored codewords by minimizing a cost function that can deal with additive noise corruptions. To compensate for linear channel distortions, the VQ-system self-adjust its codebook to the channel during online recognition. In section 3 we discuss the VQ-based approach to derive the missing data mask. The experiments on the Aurora4 large vocabulary database are presented in section 4 where we compare the VQ-masks with the harmonicity mask of [6]. Finally, conclusions can be found in section 5.

2. Missing Data Techniques

The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMM). In the front-end, a low resolution MEL-spectral representation is computed by a filter bank with D channels through windowing, framing, FFT and filter bank integration. At each frame t, let $\boldsymbol{s}_t$, $\boldsymbol{n}_t$ and $\boldsymbol{y}_t$ denote the vector of log-MEL spectral features for clean speech, noise and noisy signal respectively.

In a missing data approach, the components of $\boldsymbol{y}_t$ are divided into a reliable, $\boldsymbol{y}_{t,r}$, and an unreliable part, $\boldsymbol{y}_{t,u}$. The reliable components of $\boldsymbol{s}_t$ are approximated by $\boldsymbol{y}_{t,r}$. In the maximum likelihood per Gaussian-based imputation [2], the missing part of $\boldsymbol{s}_t$ is estimated by minimizing the (negative) log-likelihood for each Gaussian mixture component over $\boldsymbol{s}_t$:

$$\frac{1}{2}(\boldsymbol{s}_t - \boldsymbol{\mu})' \boldsymbol{P} (\boldsymbol{s}_t - \boldsymbol{\mu}) \tag{1}$$

subject to the equality and inequality constraints:

$$\boldsymbol{s}_{t,r} = \boldsymbol{y}_{t,r} \text{ and } \boldsymbol{s}_{t,u} \leq \boldsymbol{y}_{t,u} \tag{2}$$

where $\boldsymbol{\mu}$ is the Gaussian mean and $\boldsymbol{P}$ is an inverse covariance or precision matrix, both estimated on clean training data. In most MDT systems, GMMs with diagonal covariance in the

[*] This work was supported by "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

log-spectral domain are used, resulting in a diagonal structure for $\boldsymbol{P}$ and a tractable MLE for $\boldsymbol{s}$. Higher accuracies are obtained with GMMs with a diagonal covariance in the cepstral domain, in which case $\boldsymbol{P}$ becomes non-diagonal. Imputation then becomes computationally more complex since the imputation of the unreliable part now requires the solution of a *Non-Negative Least Square* problem. In this paper, the PROSPECT features defined in [2] will be used. Just like cepstra, they are computed by a linear transform of the logarithm of the filter bank energies. While these features can be applied in any speech recognition system, they show especially a clear benefit in MDT-based recognition since they reduce the computational requirements over cepstral MDT while the accuracy is maintained.

3. VQ-based Missing Features Detector

A crucial aspect in a MDT-based recognizer is the computation of the reliability masks. The ideal decision criterion labels the components of the observed noisy feature frames $\boldsymbol{y}_t$ as missing if:

$$\boldsymbol{s}_t < \boldsymbol{n}_t - \theta \tag{3}$$

where the mask threshold θ is a constant. Since clean speech and noise are unknown in real life situations, a possible mask estimation strategy is to find estimates for the log-spectra in (3). To this end, we make use of a VQ-based approach with a codebook representing the constrained space for human speech. The underlying idea is to rely on a strong model for speech, while making weak assumptions about the noise.

In vector quantization, the best codebook entry to represent the observed feature vector is selected by some perceptual distance measured between the noisy speech and the codebook vectors. Since it is the intention of the MDT-approach to avoid multistyle training, the codebook has to be trained on clean speech and mismatches the noisy input data with less accurate decoding at low SNRs as a consequence. In an attempt to compensate for this mismatch, we integrate harmonicity in the VQ-decoder. Therefore, we use the harmonic decomposition technique of [6], although alternative methods like comb filtering [7] could be used instead.

3.1. Harmonic Decomposition

To decompose a speech signal into a harmonic component and random component, a pitch estimate is first computed by a subharmonic summation method. The signal is then subsequently framed in overlapping segments with a length of two pitch periods and a single period of frame-shift. If j is the pitch epoch index and N_j the estimate of the double pitch period, then the noisy speech signal is written as:

$$y_j(n) = h_j(n) + r_j(n) \;\text{ with } 0 \leq n < N_j \tag{4}$$

with $r_j(n)$ is the random component and the harmonic component is modelled as a sum of harmonically related sinusoids:

$$h_j(n) = C_j(n) \sum_{k=0}^{K_j} a_{k,j} cos(k\omega_{0,j} n) + b_{k,j} sin(k\omega_{0,j} n) \tag{5}$$

where $\omega_{0,j}$ is the fundamental frequency estimate for segment j and K_j is total number of harmonics in the signal. The change in amplitude over the N_j samples is taken into account by the linear modulation factor $C_j(n) = (1 + c_j n/N_j)$. For each segment j, the parameters $a_{k,j}$, $b_{k,j}$, c_j and $\omega_{0,j}$ are estimated in the least square sense by minimizing the cost $(\boldsymbol{y}_j - \boldsymbol{h}_j)'(\boldsymbol{y}_j - \boldsymbol{h}_j)$. For more details we refer to [6]. The harmonic part is then

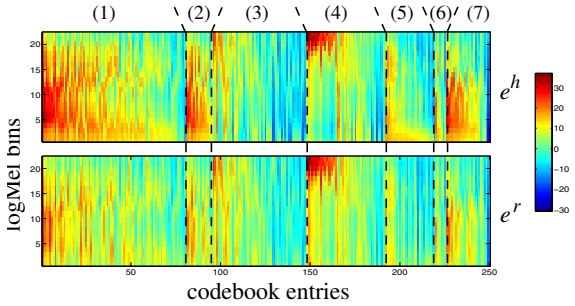

Figure 1: *VQ codebook with 250 entries representing (1) vowels, (2) dipthongs, (3) plosives, (4) (af)fricatives, (5) nasals, (6) glides and (7) liquids.*

formed by the concatenation of these segments and the random part is computed as the difference between the noisy and the harmonic signal. The log-MEL feature vectors at frame t of the harmonic and random part of a noisy speech frame $\boldsymbol{y}_t$ will be denoted by $\boldsymbol{h}_t^y$ and $\boldsymbol{r}_t^y$ respectively.

3.2. Codebook training

To create a codebook which can capture harmonic spectral information, we first apply the harmonic decomposition to the training set recorded in clean conditions. We then stack the log-Mel scaled features of the harmonic signal and those of the random part into a single vector of dimension $2D$. These vectors will be further categorized into broad phonetic classes, e.g. vowels, dipthongs, glides, liquids, (af)fricatives, nasals and stops. Frames corresponding to silence will be disregarded for reasons which become clear in 3.3. In order to provide a set of codewords representative for all phonemes, a sub-codebook was trained for each phonetic class by the k-means algorithm initialized by random vectors. In an additional refinement step, the codebook outliners were replaced by performing 2-means clustering onto the largest clusters with the highest variance. The complete codebook thus consists of seven sub-codebooks each modelling a different phoneme class. The number of codewords for a phoneme class is proportional to the occurrence percentage of this class in the training data. Table 1 shows these percentages for the training set of the Aurora4 database.

vowels	dipthongs	plosives	fricatives
32.3	6.0	21.2	16.5
affricatives	nasals	glides	liquids
1.1	11.3	2.4	9.2

Table 1: *Occurrence percentages of broad phonetic classes in Aurora4.*

The codebook vectors will be denoted by $\boldsymbol{e}_i = \left(\boldsymbol{e}_i^h, \boldsymbol{e}_i^r\right)$ where the superscripts indicate the harmonic part and random part of the codewords. Experiments have shown that no significant increase in recognition accuracy is achieved if the codebook contains more then 200 entries. A 250-entry codebook is illustrated in figure 1.

3.3. Vector Quantization of noisy speech

Since the codebook only represents a model for the human voice, decoding in non-speech (or noise) frames will lead to incorrect codebook matching and misclassifications in the mask. Therefore, a voice activity detector (VAD) based on the integrated bispectrum, inspired by [8], will segment speech from non-speech frames in order to restrict the decoding during

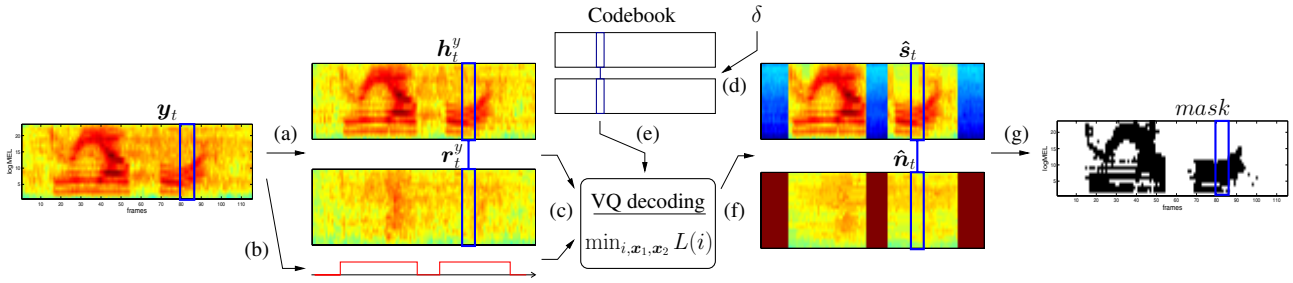

Figure 2: *Schematic overview of the VQ-based missing data detector. The noisy input signal is first decomposed into a harmonic and a random part (a) and is segmented in speech/non-speech frames by a VAD (b). During speech frames, the decoding algorithm (c) searches in the codebook for the codeword (e) that minimizes the cost $L(i)$. The codebook is updated online (d) by the channel estimate δ. After decoding, we obtain estimates for speech and noise (f) from which a missing data mask (g) is constructed.*

speech events. For a frame labeled as non-speech, all mask values will be set to one, indicating that all components are unreliable. After decomposing the incoming noisy signal $\boldsymbol{y}_t$ into harmonically related and random components as described in section 3.1, we construct the input vectors $\boldsymbol{h}_t^y$ and $\boldsymbol{r}_t^y$, for which a search is done through the entire codebook to find the best matching codeword. The proposed search criterion is to minimize the cost function

$$\min_{i,\boldsymbol{x}_1,\boldsymbol{x}_2} L(i) = \Big(\boldsymbol{h}_t^y - \max(\boldsymbol{e}_i^h, \boldsymbol{x}_1)\Big)^2 + \Big(\boldsymbol{r}_t^y - \max(\boldsymbol{e}_i^r, \boldsymbol{x}_2)\Big)^2 + \gamma\Big(\boldsymbol{x}_1 - \boldsymbol{x}_2\Big)^2 + \lambda\Big(\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_t\Big)^2 \quad (6)$$

over the unknowns $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The unknown vectors can be seen as estimates for the harmonic and random part of the noise, such that the latter two terms of (6) can easily be understood as follows. The term $(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2$ states that the harmonic and random part of the noise are close, e.g. the noise has a sufficiently smooth spectrum, and $(\boldsymbol{x}_2 - \hat{\boldsymbol{x}}_t)^2$ indicates that the random part of the noise is close to the long-term minimum of $\boldsymbol{r}^y$ as expressed in (7) below. The latter is based on the observation that the short term power spectrum of the noisy speech signal frequently decays to values which are representative of the noise power level, This observation was also used in the minimum statistics technique of [9]. Hence, the noise level can be estimated from the long-term minimum of $\boldsymbol{y}$. However, since the variance of the random part will be smaller than the variance of the harmonic part, a more accurate value for the noise level can be obtained from the minimum of $\boldsymbol{r}^y$ over a window of $2L+1$ frames (L is chosen as 10) and is defined as

$$\hat{\boldsymbol{x}}_t \approx \alpha \min_{t-L \le k \le t+L} \boldsymbol{r}_k^y \quad with\ \alpha \ge 1. \quad (7)$$

The parameters γ and λ are weights and have a fixed value.

In order to separate harmonic and random parts due to speech and noise, we have expressed through a codebook that the speech components $\boldsymbol{e}_i^h$ and $\boldsymbol{e}_i^r$ are related, while the components of the noise $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ must be similar and close to a short term sliding minimum. Because this minimum is not derived from the noisy speech itself but from its random component it is not easily perturbated by voiced speech, even at the small values of the window length parameter L that are required to track non-stationary noise.

For a given codeword, the cost is the sum of D independent cost functions, hence the computational load of the decoding can be neglected since the codebook size is small. For each frequency component, the cost is a quadratic in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, provided the max-operators do not change argument. Hence, we can minimize the cost under the assumptions $(\boldsymbol{x}_1 < \boldsymbol{e}_i^h)$ or $(\boldsymbol{x}_1 > \boldsymbol{e}_i^h)$ and $(\boldsymbol{x}_2 < \boldsymbol{e}_i^r)$ or $(\boldsymbol{x}_2 > \boldsymbol{e}_i^r)$, which gives 4 different cases, each with a different cost. However, it is possible that the minimizer violates the constraint, i.e. that minimum lies outside the region over which the cost formula is valid. In that case the true minimum lies on the boundary, e.g. $(\boldsymbol{x}_1 = \boldsymbol{e}_i^h)$. Hence, we set $\boldsymbol{x}_1$ equal to $\boldsymbol{e}_i^h$ and minimize over $\boldsymbol{x}_2$, where again we need to consider two possible solutions, $(\boldsymbol{x}_2 < \boldsymbol{e}_i^r)$ and $(\boldsymbol{x}_2 > \boldsymbol{e}_i^r)$.

By introducing the unknown parameters $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the cost function, the additive noise is taken into account. Nevertheless, there still remains a model mismatch due to channel differences between the conditions in which the codebook is trained and those during testing. Channel distortions are typically convolutional and will cause a shift $\boldsymbol{\delta}$ to the clean speech models. An algorithm for estimating this channel shift was already integrated in our MDT-based recognizer [10]. The channel is here estimated by maximizing the log-likelihood of the optimal state sequence of an observation sequence with an update time chosen as such that we have collected a sufficient amount of speech data to make a reliable estimate. In on-line applications, channel re-estimation can be postponed until the optimal state sequence becomes independent of the state in the ASR. Once a new estimate for the channel is computed, the codebook will be adjusted to the channel by subtracting $\boldsymbol{\delta}$ from all codewords:

$$\big(\boldsymbol{e}_i^h, \boldsymbol{e}_i^r\big) \leftarrow \big(\boldsymbol{e}_i^h, \boldsymbol{e}_i^r\big) - \big(\boldsymbol{\delta}, \boldsymbol{\delta}\big) \quad (8)$$

In the last step of the VQ-based missing data detector, speech and noise are reconstructed by estimating them as:

$$\begin{aligned} \hat{\boldsymbol{s}}_t &= \max(\boldsymbol{e}_i^h, \boldsymbol{e}_i^r) \\ \hat{\boldsymbol{n}}_t &= \max(\boldsymbol{x}_1, \boldsymbol{x}_2) \end{aligned} \quad (9)$$

Finally, the missing data masks are obtained by substituting these estimates in the decision criterion of (3). A schematic overview of the VQ-based missing data detector is illustrated in figure 2.

To improve the continuity of the codeword sequence, extra temporal constraints could be taken into account by training the codeword transitions in a bigram model on top of the VQ. However, experiments have shown that this does not lead to an increase in performance. This can be attributed to the fact that the bigram model also forces continuity in frames where speech transitions take place or when the energy of the voice rapidly fluctuate. As a consequence, the bigram will substitute the well-matching codewords by incorrect ones. This effect has also been discussed in [5].

Aurora4, 16 kHz sampling, clean training, multistyle testing.																	
	Close Talk								Far Talk								
TEST	1	2	3	4	5	6	7	mic 1	8	9	10	11	12	13	14	mic 2	Avg.
baseline	6.82	12.98	32.62	40.95	38.50	32.51	38.37	28.96	21.35	30.51	47.71	53.26	55.52	46.95	55.41	44.39	36.68
harm. mask	9.66	15.80	25.33	35.92	33.48	31.25	31.52	26.14	20.18	30.62	41.40	46.33	51.54	41.10	50.01	40.17	33.15
harm.+VAD mask	9.70	16.27	23.74	33.78	32.77	26.96	31.31	24.93	19.80	30.92	41.47	46.16	51.39	40.37	50.14	40.04	32.48
VQ-mask	6.33	11.66	20.55	27.42	25.46	19.41	24.73	19.37	15.26	22.85	34.19	39.06	40.95	32.51	41.01	32.26	25.81
oracle mask	6.56	8.84	13.36	15.67	16.03	11.12	16.38	12.57	14.55	19.84	27.91	30.21	32.30	25.31	32.24	26.05	18.99

Table 2: *Word error rates on Aurora4 without noise compensation, MDT with harmonicity masks, VQ-masks and oracle masks.*

4. Recognition Experiments

Experiments were conducted on the Aurora4 large vocabulary database, derived from the WSJ0 Wall Street Journal 5k-word dictation task. For each of the 2x7 test sets (no noise, car, babble, restaurant, street, airport, train), all 330 utterances from 8 different speakers, with an SNR-level that ranges from 5 dB to 15 dB, are evaluated. Test sets 01-07 were recorded with the same microphone as during training, while test sets 08-14 were recorded with a microphone selected from a set of different microphones.

The speaker-independent LVCSR-system that has been developed by the ESAT speech group of the K.U.Leuven, is used as a backend recognizer because of its fast experiment turn-around time and good baseline accuracy. The details of the MDT-recognizer can be found in [11]. In a nutshell, the 22-channel MEL filter bank spectra are transformed to the PROSPECT domain, where they are modelled with 4961 tied states in the cross-word context-dependent models and with an average of 200 Gaussians with diagonal covariance per state. A bigram language model for a 5k-word closed vocabulary is provided by Lincoln Laboratory, while decoding is done with a time-synchronous beam search algorithm.

Table 2 presents the word error rate for the MDT-based recognizer where the VQ-masks are used for the static features while leaving the dynamic features uncompensated since they are more robust to noise. These results are compared with the baseline, e.g. when no explicit noise reduction algorithm is applied and when the static mask is replaced by the harmonicity mask computed from the noisy data as described in [6] and which only exploit the harmonicity of voiced speech. As can be seen in the table, the results of the VQ-based masks are significantly better than those of the harmonicity masks. For the latter, the decision criterion uses the idea that the harmonic part will be dominated by the speech which may lead to poor decisions in unvoiced speech segments. To illustrate that the VQ-model most contributes to the improvement and not the use of a VAD in the missing data detector, we have added an extra experiment where the harmonicity mask is used during non-speech frames and all mask values are set equal to one during non-speech frames as is done for the VQ-mask. For separating unvoiced speech from background noise, the VQ-based approach also relies on the constraints of the spectral shape captured by the codebook. Further potential of the MDT-approach is shown by applying oracle masks where (3) is used as a decision criterion.

5. Discussion and Conclusions

The detection of missing data often relies on important assumptions about the noise type and will consequently fail in environments that do not meet these assumptions. Therefore we aim to exploit more refined models of the human voice while exploiting noise knowledge minimally. To this end, we proposed a VQ-based approach where speech characteristics as harmonicity, voicing, spectral shape (timbre) and to some extent onsets serve to obtain better masks. Experiments conducted on the Aurora4 database have shown that this approach can deal with a broad set of noise types and microphone characteristics and show uniformly better accuracy than previously reported.

6. References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Comm.*, vol. 34, 2001, pp. 267–285.

[2] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, October 2004, pp. 101–104.

[3] A. Gersho and R. Gray, "Vector quantization and signal compression." Kluwer Academic Press, 1992.

[4] S. So and K. K. Paliwal, "Improved noise-robustness in distributed speech recognition via perceptually-weighted vector quantisation of filterbank energies," in *Proc. ICSLP*, Lisbon, Portugal, Sept. 2005.

[5] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. ICASSP*, Toulouse, France, 2006.

[6] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 213–216.

[7] B. R. M.L. Seltzer and R. Stern, "Classifier-based mask estimation for missing feature methods of robust speech recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[8] J. Ramírez, J. Górriz, J. Segura, C. Puntonet, and A.Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum lrt," in *IEEE Signal Processing Letters*, 2006.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, July 2001, pp. 504–512.

[10] M. Van Segbroeck and H. Van hamme, "Handling convolutional noise in missing data automatic speech recognition," in *Proc. ICSLP*, Pittsburgh, U.S.A., Sept. 2006, pp. 2526–2565.

[11] ESAT-PSI Speech Group website, http://www.esat.kuleuven.be/psi/spraak/.