IDIAP RESEARCH REPORT

LP-TRAPs in all senses

Petr Motlicek *

IDIAP–RR 07-66

December 2007

* IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**

Av. des Prés-Beudin 20 P.O. Box 592 1920 Martigny – Switzerland

Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 07-66

LP-TRAPs in all senses

Petr Motlicek

December 2007

Résumé. This report describes additional experiments with LP-TRAPs – speech features derived from autoregressive model applied to approximate temporal evolution of speech spectra in critical band-sized frequency sub-bands. The importance of free parameters, such as order model, length of the approximated temporal pattern, compression factor, or number of resulting cepstral coefficients, is investigated and evaluated using TANDEM-ASR approach.

1 Introduction

Traditional speech feature extraction segments input signal into short-term ($\sim$ 20 ms long) frames. The resulting spectrogram which represents temporal evolution of spectral parameters of the short-term frames is therefore largely downsampled ($\sim$100 Hz) in contrast to the speech signal. Similarly long-term speech feature extraction techniques, such as TempoRAl Patterns (TRAPs) [1] as well as recently developed M-RASTA [2] technique, derive speech parameters from the short-term spectral samples. Therefore, these techniques cannot imply any finer temporal details. In other words, extracted sub-band speech features capture information from long temporal context with low temporal resolution.

Furthermore, TRAPs represent long-term energies of the sub-band signal, thus directly parameters of the signal. It would be desirable to work rather with parameters of a model, modeling these temporal vectors. This is a case in e.g., short-term frame based PLP feature extraction technique, where power spectra of the short-term speech frames are approximated by AR model [3]. In addition, order of the model can be used as a parameter to control smoothness, i.e., how well the model approximates original trajectory [4].

One of the speech parameterization which is able to capture finer temporal details of the speech signal can be based on Hilbert envelope estimated in each frequency sub-band. Although, Hilbert envelope represents an energy trajectory with full temporal resolution, it needs to be smoothed to suppress undesirable information from ASR point of view, such as the one captured by glottal pulses. In principle, this is a case of Linear Predictive TRAP (LP-TRAP) feature extraction technique, where smoothing is done using Auto-Regressive (AR) model. LP-TRAPs, first proposed in [5, 6], apply AR model in a dual way to PLP technique, to model the sub-band temporal trajectory while preserving full temporal resolution. In other words, parameters of AR model describe smoothed version of the sub-band trajectory (temporal envelope). Profound theoretical background can be found in [7].

The goal of this report is to evaluate LP-TRAP technique according to its sensitivity to free parameters. Unlike [6], we also take into account length of temporal segments and number of cepstral coefficients together with model order and compression factor.

2 Estimation of temporal envelopes in frequency sub-band

Traditionally, AR modeling has been used in speech feature extraction to parameterize spectral envelopes of short-term frames (referred to as Temporal Domain Linear Prediction (TDLP) approach). Resulting LP coefficients are estimated in such a way that error between signal power spectrum $P(\omega_k)$ (sampled at discrete frequencies $\omega_k = \frac{2\pi}{N}k\,;\ k = 1, \ldots, N\,;$ in PLP technique obtained using critical sub-band decomposition and the compression) and model power spectrum $\hat{P}(\omega_k)$ is minimized [8] :

$$E_{TDLP} \approx \frac{1}{N}\sum_{k=1}^{N}\frac{P(\omega_k)}{\hat{P}(\omega_k)}, \tag{1}$$

where N denotes length of the input time-segment corresponding to P.

LP-TRAPs use AR model to approximate temporal envelope of the sub-band signals (referred to as Frequency Domain Linear Prediction (FDLP) approach). In order to obtain a model approximating temporal envelope of the sub-band signal, coefficients of AR model need to be estimated in frequency domain. This can be done by performing DCT on full-band time-domain signal followed by segmentation of DCT sequence into frequency sub-bands. In particular, the segmentation is performed by applying set of Gaussian windows of variable temporal resolution, spaced following the Bark scale (depicted in Figure 1). This results in DCT sequences where each sequence represents critically band-sized frequency sub-band. These DCT sequences can then be used to estimate sub-band AR models. It can be shown (e.g. [9]) that prediction error is then proportional to the integrated ratio of

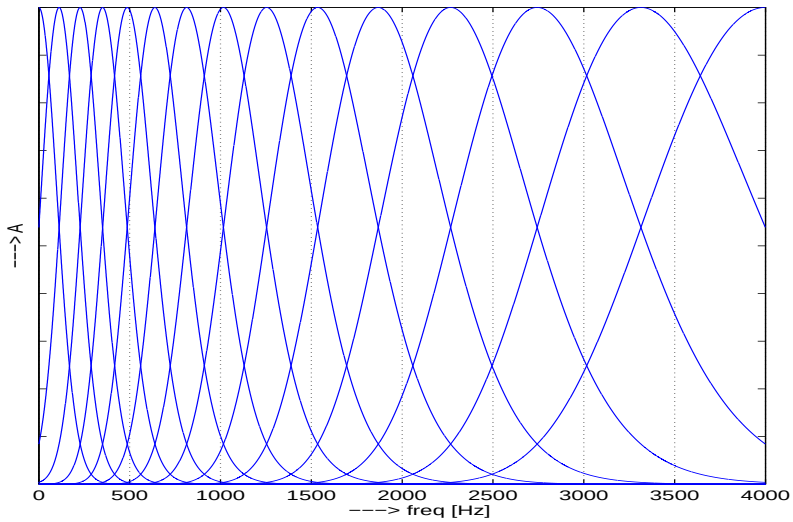

FIG. 1 – *Set of Gaussian windows employed to perform frequency decomposition into critically band-sized frequency sub-bands.*

the squared Hilbert envelope $|\phi(m)|^2$ and its AR model approximation $A(m)^2$:

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^2}{A^2(m)}. \tag{2}$$

$\phi(m)$ is an analytic signal obtained from sub-band signal (the magnitude of $\phi(m)$ represents Hilbert envelope in a given frequency sub-band). Eq. 2 can be interpreted in such a way that the FDLP AR model fits squared Hilbert envelope of the symmetrically extended time-domain signal. FDLP models the time-domain envelope in the same way as TDLP models the spectral envelope.

In addition, the squared Hilbert envelope $|\phi(m)|^2$ is available and can be modified. Thus, e.g., compressing $|\phi(m)|^2$ by a root function $[.]^{\frac{1}{r}}$ turns Eq. 2 into :

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^{\frac{2}{r}}}{A^{\frac{2}{r}}(m)}. \tag{3}$$

As a consequence, by modifying r we can control the fit of peaks as well as dips of the original Hilbert envelope. This technique has been proposed for speech feature extraction based on TDLP and can be also successfully applied in LP-TRAPs.

3 Experiments

3.1 Experimental setup

The front-end algorithm for speech recognition is evaluated using TANDEM-ASR approach [10]. Derived LP-TRAPs are first fed to MLP which is discriminatively trained to classify phonemes. In particular, set of MLPs is trained in each frequency sub-band independently (denoted as Band-MLPs) and their outputs are then merged using another MLP (Merger-MLP). Input features transformed by 2-layer MLP architecture are then fed into a conventional GMM-HMM recognizer. The speech databases (DBs) OGI Stories [11] and OGI Numbers95 [12] were used, which consist of 8 kHz speech

recorded over a telephone channel. OGI Stories DB contains spontaneous continuous speech, while OGI Numbers95 DB contains recordings of strings of digits and numbers.

Band-MLPs were trained on 2.8 hours of OGI Stories (208 files). Merger-MLP was trained on 1.7 hours of OGI Numbers95 (3590 files). All MLP classifiers were discriminatively trained to classify 29 phonemes. 90% of the data was used for training, 10% for cross-validation.

GMM-HMM recognizer was trained on 1.3 hours (2547 files - subset of the set for Merger-MLP), and tested on 1.7 hours (2169 files) of OGI Numebrs95 (strings of 11 digits with word transcriptions). We used 22 phoneme HMMs (out of 29 from MLPs), 5 emitting states, 32 Gaussian mixtures, 11 target words in 28 pronunciation variants.

3.2 MLP setup

Set of LP-TRAP features was extracted for each critically band-sized frequency sub-band. Band-MLP classifier has three layers : $N \times 100 \times 29$ (N features at the input, 100 hidden units, 29 phoneme outputs). Each Band-MLP was trained using approximately 1000k input vectors (approximately 1/10 was used for MLP cross-validation to determine the end of training). Phoneme posteriors from all critical sub-bands are merged by Merger-MLP : $435 \times 300 \times 29$ (15 Band-MLP$\times$29 features, 300 hidden units, 29 phoneme outputs). Approximately 600k input vectors were used to train Merger-MLP.

3.3 Evaluation criteria

Word Error Rate (WER) measure obtained on the test set of OGI Numbers95 (strings of 11 digits) is used to evaluate LP-TRAPs and to compare them with other (state-of-the-art) feature extraction techniques. Evaluation of free parameters in LP-TRAPs technique is done using relative WER with respect to the "baseline" LP-TRAP parameter (experimental) setting :

$$rel.\ WER = \frac{W_{BAS} - W}{W_{BAS}}, \tag{4}$$

where W_{BAS} denotes absolute WER of baseline setting and W denotes WER of the evaluated setting.

3.4 LP-TRAPs – free parameters

LP-TRAPs have various free parameters, which play an important role and need to be experimentally estimated to acquire good ASR performances. Some experimental work has already been done and is presented in [5, 6]. Among the most important parameters belong : **(a)** length of the temporal window, **(b)** order of AR model, **(c)** compression factor, and **(d)** number of cepstral coefficients recursively computed from obtained AR model.

Furthermore, we consider two types of LP-TRAP features that can be extracted from AR polynomials. In the first approach, referred to as **ENV**, temporal envelope is extracted by sampling frequency response of the resulting AR model. In the second approach, referred to as **CEPS**, the cepstral recursion is performed which allows for converting AR model of the temporal trajectory into modulation spectra with arbitrary number of cepstral coefficients.

LP-TRAP parameters in details :

(a) *Length of temporal window :* This parameter determines the length of temporal trajectory applied on the input signal to compute DCT coefficients and in consequence to estimate AR model. Temporal window is applied with 10ms shift so that we obtain set of LP-TRAP features with sampling frequency 100 Hz (it does not mean that resulting temporal envelopes are extracted in down-sampled domain). We analyzed two window lengths, 1000 and 500 ms. Pilot experiments were done with 1000 ms long temporal windows.

(b) *Order of AR model - Nb : Nb* varied from very low to relatively high values. However, estimated LP coefficients were not directly used as input features for TANDEM-ASR. In case of **ENV** approach, magnitude frequency response sampled with (101 or 51 points, depending on the

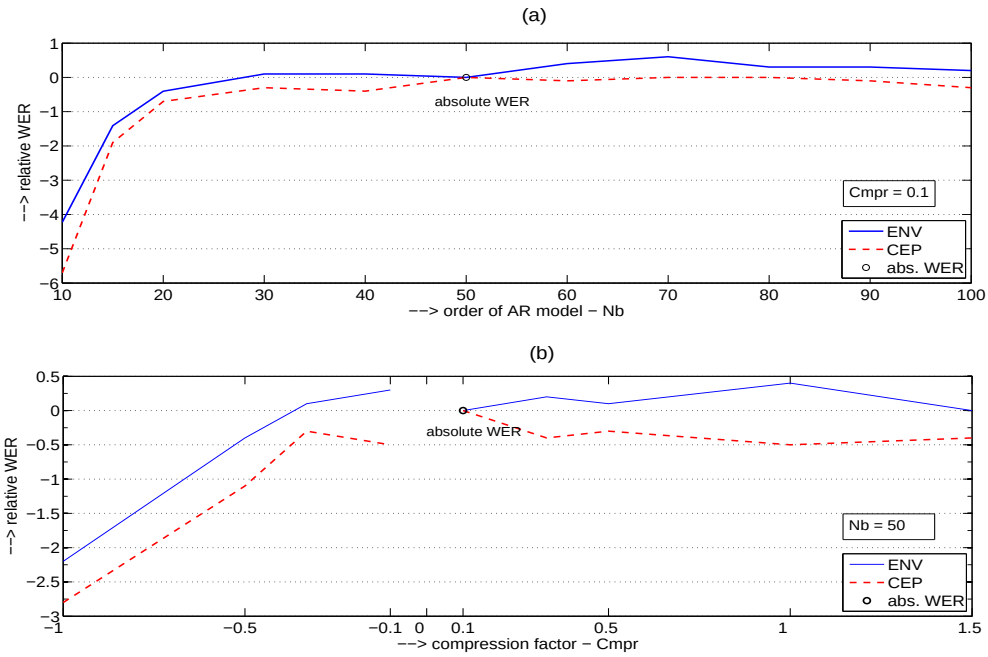

FIG. 2 – *Plot of relative WER as a function of : (a) order of AR model, (b) compression factor applied to derive AR model. Length of temporal trajectories is* 1000 *ms. Compression ratio equal to* 0 *is obviously excluded, since such AR model would lead to the flat magnitude frequency response. Absolute WER (W_{BAS}) for ENV is equal to* 4.9%*. Absolute WER for CEP is equal to* 4.6%*.*

length of temporal window) provides output features. In case of **CEP** approach, 51 cepstral coefficients are derived to be used as speech features for TANDEM-ASR. Experimental results are given in Tables 2 and 4.

(c) *Compression factor - Cmpr :* This parameter, denoted as r in Eq. 3, determines the ratio between approximating peaks and dips of the temporal envelope by the model. AR model for $Cmpr = 0$ is not defined since it would lead to the model with constant magnitude frequency response. Experimental results are given in Tables 3 and 5.

(d) *Number of cepstral coefficients - Cps :* In case of 500 ms LP-TRAPs, we also experimented with number of cepstral coefficients (thus only related to **CEP** approach) to be estimated from AR model. The energy term c_0 is always excluded. Experimental results are given in Table 6.

3.5 Experimental results

With regards to the experimental setup, the best absolute WER score obtained with LP-TRAP speech features is 4.1%. This performance is compared to state-of-the-art speech feature extraction systems processing long temporal context - TRAPs (101 points), M-RASTA ($\Delta f + \Delta^2 f$) and PLP-TANDEM, with experimental results given in Table 1. We also added performances of traditional short-term frame based PLP features. LP-TRAPs (the best system) are also combined with M-RASTA technique (frequency response of AR model is used as the input to estimate M-RASTA features).

Experiments with fixed compression factor and varying order of AR model show that continuous increase of the order results in an increased performances until the saturation point. When 1000 ms long temporal segments are considered, the recognition improvement starts saturating for $Nb \sim (50 - 70)$, as seen in Figure 2(a). Similar behavior can be found for 500 ms long temporal segments for $Nb \sim 30$

system	WER [%]
TRAPs	4.7[a]
M-RASTA	3.6
PLP	5.2[b]
PLP-TANDEM	4.8[c]
LP-TRAPs	4.1
LP-TRAPs + M-RASTA	3.8

[a] 1000 ms long temporal patterns, 101 inputs for 15 Band-MLPs.
[b] AR model order= 15, 12 cepstral coefficients + c_0 + Δ + $\Delta\Delta$, no feature transform by MLP.
[c] 9 consecutive frames of the PLP features (351 dimensions), one MLP trained.

TAB. 1 – *The best performances (absolute WER) achieved with LP-TRAPs compared to the state-of-the-art systems. Baseline experimental results for TRAPs, PLP and PLP-TANDEM are obtained from the PhD thesis of Petr Fousek, Czech Technical University of Prague, 2007.*

(refer to Table 4).

Following experimental results with fixed order of AR model and varying compression factor, depicted in Figure 2(b), indicate that better performances can be achieved with $Cmpr \sim (0, 1\rangle$. This means that LP-TRAPs prefer more accurate modeling not only peaks but also dips of the sub-band temporal envelopes.

In case of 500 ms LP-TRAPs, number of cepstral parameters derived from estimated AR model varied. For these experiments, we also modified number of hidden units in Band-MLPs (lesser for smaller number of inputs). Less cepstral coefficients derived from AR model did not have any large impact on the achieved performaces, as can be seen from Table 6.

4 Discussion and Conclusions

By using AR model for modeling sub-band temporal envelopes, we can extract temporal structure of the speech without need for short-term frame processing. Resulting LP-TRAP speech parameters expressed in terms of smoothed temporal patterns or cepstral features derived from AR model were evaluated using TANDEM-ASR approach. The best achieved WERs outperform traditional TRAP and PLP based features and perform slightly worse than recently proposed M-RASTA feature extraction.

The main goal of this report was to present performances of LP-TRAPs for various parameter setting. In general, the best performances were obtained for 500 ms long LP-TRAPs (cepstral representation) approximated by AR model $Nb = 50, Cmpr = 0.1$.

5 Acknowledgments

This work was supported the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2" ; and managed by the IDIAP Research Institute on behalf of the Swiss Federal Authorities, and by the European Commission 6^{th} Framework DIRAC Integrated Project.

Références

[1] H. Hermansky, S. Sharma. "Temporal patterns (TRAPs) in ASR of noisy speech", *in Proc. ICASSP*, vol. 1, pp. 289-292, March 1999.

[2] H. Hermansky, P. Fousek. "Multi-resolution RASTA filtering for TANDEM-based ASR", *in Proc. Interspeech 2005*, pp. 361-364, Lisbon, Portugal, September 2005.

[3] H. Hermansky. "Perceptual linear predictive (PLP) analysis for speech", *J. Acoust. Soc. Am.*, pp. 1738-1752, 1990.

[4] H. Hermansky, H. Fujisaki, Y. Sato. "Analysis and Synthesis of Speech based on Spectral Transform Linear Predictive Method", *in Proc. of ICASSP*, Vol. 8, pp. 777-780, Boston, USA, April 1983.

[5] M. Athineos, D. Ellis, "Frequency-domain linear prediction for temporal features", *Automatic Speech Recognition and Understanding Workshop IEEE ASRU* , pp. 261-266, December 2003.

[6] M. Athineos, H. Hermansky, D. Ellis. "LP-TRAP : Linear predictive temporal patterns", *in Proc. of ICSLP*, pp. 1154-1157, Jeju, S. Korea, October 2004.

[7] R. Kumaresan, A. Rao. "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", em J. Acoust. Soc. Am., 105 (3), pp. 1912 - 1924, March 1999.

[8] J. Makhoul. "Linear Prediction : A Tutorial Review", *in Proc. of IEEE*, Vol. 63, No. 4, April 1975.

[9] P. Motlicek, V. Ullal, H. Hermansky, "Wide-Band Perceptual Audio Coding based on Frequency-domain Linear Prediction", *in Proc. of ICASSP*, Honolulu, USA, April 2007.

[10] H. Hermansky, D. Ellis, S. Sharma. "Tandem connectionist feature stream extraction for conventional HMM systems", *in Proc. of ICASSP*, Istanbul, Turkey, 2000.

[11] R. Cole, R. Noel, T. Lander. "Telephone speech corpus development at CSLU", *in Proc. ICSLP'94*, pp. 1815-1818, Yokohama, Japan, 1994.

[12] R. Cole, R. Noel, T. Lander, T. Durham. " New telephone speech corpora at CSLU", *in Proc. Fourth European Conference on Speech Communication and Technology*, vol. 1, pp. 821-824, 1995.

	Nb :	10	15	20	30	40	50	60	70	80	90	100	200
ENV :	**Rel. WER :**	-4.2	-1.4	-0.4	0.1	0.1	0 **(4.9)**	0.4	0.6	0.3	0.3	0.2	0.1
CEP :	**Rel. WER :**	-5.7	-1.9	-0.7	-0.3	-0.4	0 **(4.6)**	-0.1	0	0	-0.1	-0.3	-0.3

TAB. 2 – *Relative WERs of LP-TRAPs for different order model Nb (for fixed $Cmpr = 0.1$). Relative error rates are associated with absolute WER (W_{BAS}) for $Nb = 50$, shown as bold text. Length of temporal trajectories is* 1000 *ms. ENV refers to temporal envelope (of length* 101*) based features. CEPS refers to* 51 *cepstral features derived from LP-TRAP.*

	Cmpr :	-1.0	-0.5	-0.33	-0.1	0.1	0.33	0.5	1.0	1.5
ENV :	**Rel. WER :**	-2.2	-0.4	0.1	0.3	0 **(4.9)**	0.2	0.1	0.4	0
CEP :	**Rel. WER :**	-2.8	-1.1	-0.3	-0.5	0 **(4.6)**	-0.4	-0.3	-0.5	-0.4

TAB. 3 – *Relative WERs of LP-TRAPs for different compression factor $Cmpr$ (for fixed $Nb = 50$). Relative error rates are associated with absolute WER (W_{BAS}) for $Cmpr = 0.1$, shown as bold text. Length of temporal trajectories is* 1000 *ms. ENV refers to temporal envelope (of length* 101*) based features. CEPS refers to* 51 *cepstral features derived from LP-TRAP.*

	Nb :	5	10	20	30	40	50
ENV :	**Rel. WER** :	-4.7	-0.5	-0.1	0	-0.2	0 **(4.9)**
CEP :	**Rel. WER** :	-6.3	-1.7	-0.3	-0.5	-0.2	0 **(4.4)**

TAB. 4 – *Relative WERs of LP-TRAPs for different order model Nb (for fixed $Cmpr = 0.1$). Relative error rates are associated with absolute WER (W_{BAS}) for $Nb = 50$, shown as bold text. Length of temporal trajectories is* 500 *ms. ENV refers to temporal envelope (of length* 51*) based features. CEPS refers to* 51 *cepstral features derived from LP-TRAP.*

	Cmpr :	0.1	0.5	1.0
ENV :	**Rel. WER** :	0 **(4.9)**	-0.2	0.1
CEP :	**Rel. WER** :	0 **(4.4)**	-0.3	0

TAB. 5 – *Relative WERs of LP-TRAPs for different compression factor $Cmpr$ (for fixed $Nb = 50$). Relative error rates are associated with absolute WER (W_{BAS}) for $Cmpr = 0.1$, shown as bold text. Length of temporal trajectories is* 500 *ms. ENV refers to temporal envelope (of length* 51*) based features. CEPS refers to* 51 *cepstral features derived from LP-TRAP.*

	Cps :	16	26	51	76	101
CEP :	**Rel. WER** :	-0.1	-0.2	0 **(4.4)**	0	0.3

TAB. 6 – *Relative WERs of LP-TRAPs for different number of cepstral coefficients Cps derived from LP model. Relative error rates are associated with absolute WER (W_{BAS}) for $Cps = 51$, shown as bold text. Length of temporal trajectories is* 500 *ms.*