EVALUATION OF PNCC AND EXTENDED SPECTRAL SUBTRACTION METHODS FOR ROBUST SPEECH RECOGNITION

Thibaut Fux[*,†,‡], *Denis Jouvet*[*,†,‡]

[*] Inria, 615 rue du Jardin Botanique, F-54600, Villers-lès-Nancy, France
[†]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[‡]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France

ABSTRACT

This paper evaluates the robustness of different approaches for speech recognition with respect to signal-to-noise ratio (SNR), to signal level and to presence of non-speech data before and after utterances to be recognized. Three types of noise robust features are considered: Power Normalized Cepstral Coefficients (PNCC), Mel-Frequency Cepstral Coefficients (MFCC) after applying an extended spectral subtraction method, and Sphinx embedded denoising features from recent sphinx versions. Although removing C0 in MFCC-based features leads to a slight decrease in speech recognition performance, it makes the speech recognition system independent on the speech signal level. With multi-condition training, the three sets of noise-robust features lead to a rather similar behavior of performance with respect to SNR and presence of non-speech data. Overall, best performance is achieved with the extended spectral subtraction approach. Also, the performance of the PNCC features appears to be dependent on the initialization of the normalization factor.

Index Terms— Speech recognition, Speech level robustness, Noise robustness, Spectral subtraction, PNCC.

1. INTRODUCTION

Speech recognition in noisy environment is one of the most studied topics over the last years. Indeed, if clean speech recognition systems are regularly improved, one of the biggest difficulty is the robustness of such systems in real conditions. Real conditions often involve background noise, short sentences and non-speech data (silence or noise) before and after the actual utterance to be recognized. Additionally, the input speech signal can also vary a lot in term of level due to the speech production level and to the distance between the user and the microphone.

All these "perturbations" degrade the speech recognition performance. The most problematic perturbation is certainly the background noise which has drawn a lot of attention from researchers since many years (see [1] for a review). Two kinds of approaches are especially developed to improve speech recognition performance in noisy environment. The first one consists to enhance the speech signal before recognizing it. This corresponds to a filtering procedure which aims to remove as much as possible the background noise [2]. The second one consists in using features, mainly inspired by human auditory processing, which are more robust to the noise than the conventional Mel-Frequency Cepstral Coefficients (MFCC) [1].

Spectral subtraction [3] is a well known noise reduction technique. This technique aims at removing the background noise (i.e., an additive noise) by subtracting an estimation of the noise spectrum from the noisy speech spectrum. Most of the time, a Voice Activity Detector (VAD) is needed to detect non-speech regions in order to update the estimation of the noise spectrum. Noise robust features are inspired by human auditory processing as for example Perceptual Linear Predictive (PLP) [4], Relative Spectral Transform-PLP (RASTA-PLP) [5] and Power Normalized Cepstral Coefficients PNCC [6]. They include some psychoacoustic effects such as temporal masking effect, and different filter bank distributions and filter shapes to better match the human audition process. In this paper we are particular interested on the recent PNCC features since they correspond to an emerging and certainly promising approach.

The previous mentioned methods (speech enhancement and robust features) are all efficient to improve speech recognition performance in noisy conditions compared to standard MFCC front end. However, their performance can be affected by the input signal level. In this paper, a solution based on the removal of the C_0 coefficient which is related to the energy, is evaluated. The impact of removing the C_0 coefficient on speech recognition performance in presence of non-speech data (i.e., silence or noise segments) is also analyzed.

In this paper, the objective is to evaluate and compare the performance of a few sets of noise robust features, namely *PNCC*, *extended spectral subtraction* and *Sphinx embedded denoising* methods. For performance evaluations, speech data utterances are extracted from French broadcast record-

The work presented in this article is part of the RAPSODIE project, and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER) (http://erocca.com/rapsodie). This work has also been partly realized thanks to the support of the "Région Lorraine" and the CPER MISN TALC project.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

ing. These segments are then altered in different ways to simulate operational conditions: by adding non-speech, by adding longer non-speech segments to the actual speech segments (i.e. resulting in non-speech regions before and after speech segments), and by attenuating the input signal level (i.e., the signal amplitude).

The paper is organized as follows. Section 2 details the three selected approaches. Section 3 describes the recognition models as well as the training and test data. Section 4 focuses on the results and Section 5 concludes the paper.

2. SPEECH ENHANCEMENT AND NOISE ROBUST FEATURES

Three single channel approaches are considered in this study: the PNCC features, the extended spectral subtraction method and the embedded Sphinx denoising method.

2.1. Power normalized cepstral coefficients (PNCC)

PNCC features were proposed by Kim and Stern [6] and are based on human auditory processing. PNCC differ from MFCC in several aspects. First, the traditional logarithmic nonlinearity used in MFCC computation is replaced by a power-law nonlinearity. A similar nonlinear function was used for computing RASTA-PLP coefficients, however for PNCC the power value is set to 1/15 according to psychoacoustic observations. Secondly, the triangular filter bank is replaced by a gammatone filter bank. PNCC computation also includes a noise suppression algorithm based on asymmetric filtering including temporal masking effects. Since PNCC computation does not use a logarithmic nonlinearity, all the features may strongly be influenced by the signal level. In order to reduce this phenomenon, the PNCC algorithm includes a power normalization procedure which consists in scaling the power of each frame according to a normalization factor estimated on-line using the power of past frames.

2.2. Extended spectral subtraction

The extended spectral subtraction method was proposed by Sovka et al. [7]. The main advantage of this approach is the combination of a Wiener filter and a spectral subtraction method. Furthermore, there is no need for a VAD since the algorithm estimates the noise pattern during speech sequences. The approach consists in using a Wiener filter to estimate the spectral noise pattern which is then subtracted from the input noisy speech spectrum. The noise spectrum is estimated on-line based on the difference between the changing rate of noise and of speech (i.e., the speech signal is assumed to change faster than the background noise). The particularity of this method is the updating of the Wiener filter using its output instead of the input, as it is usually done when a Wiener filter is used (see [7] for more details). In addition to the spectral subtraction, corrections are added through a half-wave rectifier, in order to reduce spectral error, and a noise overestimation factor (see [8] for more details).

2.3. Sphinx embedded denoising features

Recent versions of the Sphinx toolkit include a denoising algorithm largely inspired from the PNCC approach. Indeed, the main difference between the Sphinx implementation and the original PNCC implementation is that, in the Sphinx version, the filter bank is the triangular one traditionally used in MFCC estimation. Furthermore, in the Sphinx version the logarithm function is still used after noise removal instead of the power-law used in PNCC. Thus in the Sphinx version there is no need to apply power normalization as in PNCC. In other words, the Sphinx version uses the asymmetric noise suppression procedure of the PNCC algorithm but the pre-processing and the post-processing correspond to that of the MFCC computation.

3. EXPERIMENTAL SETUP

This section describes the data used to train the speech recognition models and the noise data collected in a noisy environment. Then, details about the acoustic analyses and the acoustic models are given.

3.1. Speech corpora

The speech corpora used in our experiments come from the ESTER2 [9] and the ETAPE [10] evaluation campaigns, and the EPAC [11] project. The ESTER2 and EPAC data are French broadcast news collected from various radio channels, thus they contain prepared speech, plus interviews. A large part of the speech data is of studio quality, and some parts are of telephone quality. On the opposite, the ETAPE data correspond to debates collected from various radio and TV channels. Thus this is mainly spontaneous speech. The speech data of the ESTER2 and ETAPE train sets, as well as the transcribed data from EPAC corpus, were used to train the acoustic models. The training data amounts to almost 300 hours of signal and almost 4 million running words. The test data corresponds to a single broadcast recording from the ESTER2 corpus (not used in the training step) which has been cut in small segments (n = 246) of various length according to the transcription file. The length of the segments (or utterances) varies from 0,7 second to 16 seconds.

3.2. Noise data

The noise used in our experiments was recorded in a shop near a cash register. Thus, the noise is a mixture of background music, background speech and cash register beeps. A three hour recording was cut into two parts: one part is used in the training step and the other part is used in the decoding step. To

generate noisy data, for each speech segment, a noise segment is randomly extracted and added to the speech signal.

Because cash register beeps, but also some impacts on the recording material, induce strong peaks in the signal, the noise signal was pre-processed to flatten the amplitude envelop. This was done by equalizing the intensity frame by frame to a reference value before reconstructing the signal. For this processing, frames of 20 ms duration, with 50% overlap, are used.

3.3. Acoustic analysis

Features extraction for both training and decoding were computed as follow:

- *MFCC with* C_0 features are extracted using Sphinx3 tools [12] [1]. Extraction is performed using 40 filters, 25.6 ms frame length, 100 frames per second and 39 coefficients (13 static from C_0 to C_{12} + 13 Δ + 13 $\Delta\Delta$). These features are used for the baseline model.
- *MFCC without* C_0 (here after denoted noC_0) features are similar to the previous ones, except that the coefficient C_0 is removed. Thus, only 38 coefficients are retained. This set of 38 features features is used for the baseline model without C_0, for Extended spectral subtraction approach and for the sphinx embedded denoising approach.
- *PNCC* features are extracted using the free C library[2] developed by Principi et al. [13]. The analysis frame length, the number of frames per second and the number of coefficients are the same as for the MFCC analysis. They are used for evaluating the PNCC approach.

3.4. Modelizations

The Sphinx3 tools [12] are used to train the context dependent phone acoustic models and to decode audio signals. The acoustic HMM are modeled with 64-Gaussian components per mixture. Five models are trained using the database described in section 3.1. Two of them are the baseline models and are trained using MFCC without any pre-pocessing, with and without the C_0 coefficient, using clean speech corpora. These two models aim to analyze the influence of the C_0 parameter. To evaluate the 3 noise robust approaches, 3 multi-condition models are trained, one for each approach. The SNR considered for the multi-condition training are: SNR = {Inf; 20 dB; 15 dB; 10 dB; 5 dB}. Noisy data are obtained by adding randomly selected noise segments to each speech segment of the training set. Before being added, the noise signal is scaled according to the required SNR value. Details of these five models are given below.

- *Baseline*: single-condition training, clean speech, MFCC with C_0.

[1]The version used in this article is the Sphinx3 R-12729, http://sourceforge.net/p/cmusphinx/code/12729/tree/trunk/

[2]url: http://a3lab.dibet.univpm.it/projects/libpncc

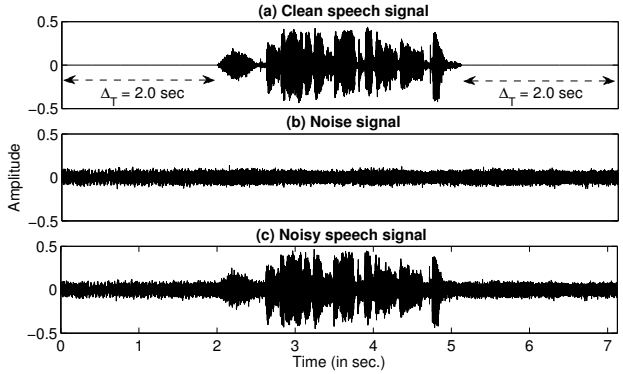

Fig. 1. Waveforms of (a) a clean speech signal utterance with silence segments of length Δ_T = 2.0 sec. before and after, (b) a noise signal (randomly extracted from the noise data) and (c) the sum of (a) and (b) which provides the noisy segment (here SNR = 10 dB)

- *Baseline_no*C_0: single-condition training, clean speech, MFCC without C_0.
- *Extended_SS_multi_no*C_0: multi-condition training, noisy speech denoised using the extended spectral subtraction method, MFCC without C_0.
- *SphinxDenoise_multi_no*C_0: multi-condition training, noisy speech denoised using Sphinx embedded method, MFCC without C_0.
- *PNCC_multi*: multi-condition training, noisy speech, PNCC features.

3.5. Tests data

Three kinds of test sets are generated based on short segments: (1) level attenuated utterances, (2) noisy utterances, (3) enlarged noisy segments (i.e., with non-speech regions before and after the actual utterances).

1. Level attenuated utterances are obtained by lowering the signal level. Four levels of amplitude attenuation are used: Att. = {0 dB; -12 dB; -24 dB; -36 dB}. Note that these attenuations correspond to dividing the signal sample values by 1, 4, 16, and 64 respectively.
2. Noisy data are generated by adding noise. For each utterance the level of the noise signal is set in order to obtain 4 levels of signal-to-noise ratio: SNR = {Inf; 20 dB; 15 dB; 10 dB}. Note that SNR=Inf corresponds to the clean speech.
3. To evaluate the influence of non-speech segments before and after utterances, silence is added at the beginning and at the end of each utterance. The duration of these segments are the following : Δ_T = {0 s; 0.5 s; 1.0 s; 2.0 s; 4.0 s}. Note that Δ_T = 1.0 s means that one second is added over the whole segment. After that, the noise is added according to three SNR levels : SNR = {20 dB; 15 dB; 10 dB}. An example is given in Figure 1.

Models \ Att.	0 dB	-12 dB	-24 dB	-36 dB
Baseline	31.53	31.06	55.61	94.67
Baseline_noC_0	32.74	32.77	33.22	33.47
Extended_SS_multi_noC_0	30.79	31.00	30.81	30.11
SphinxDenoise_multi_noC_0	32.55	31.73	32.53	32.56
PNCC_multi	33.29	42.92	93.60	94.94
PNCC_multi+Adapted_ μ_0[3]	*33.29*	*33.53*	*33.14*	*32.61*

Table 1. WER (%) according to the attenuation of the input signal (Att.).

4. RESULTS

After evaluating the impact of the C_0 coefficient, this section analyzes the performance of the sets of noise robust features.

4.1. Input level robustness

The impact of the input signal level on speech recognition performance is analyzed using clean speech data (SNR = Inf) and various attenuation values (Att = {0 dB; -12 dB; -24 dB; -36 dB}), but without adding non-speech segments (Δ_T = 0 s).

Table 1 gives the Word Error Rate (WER) for the various sets of features according to the attenuation level. It is clear that the *Baseline* model (i.e., with C_0) is very sensitive to this attenuation and the performance strongly degrades for large attenuation values (i.e., -24 dB and -36 dB). However, removing the C_0 coefficient (i.e., *Baseline_noC_0*) gives very good results, even for large attenuation values (i.e., -36 dB), and the performance just slightly decreases. This justifies our motivation to remove the C_0 coefficient in all sets of MFCC-based features in order to improve the robustness with respect to signal level. Evaluations done with the other MFCC-based features used in this paper give similar results when the C_0 coefficient is removed.

Concerning the noise robust feature sets, because two of them are based on MFCC without C_0, their performance are comparable and are not sensitive to the input level (see Table 1). But, for the model trained using PNCC, features the WER strongly increases even with a low attenuation value such as -12 dB. This point is discussed in section 4.4.

4.2. Noise robustness

Table 2 shows the WER results according to SNR. Concerning baseline features, a slight degradation can be observed for the baseline model without C_0, compared to the model including C_0, for clean speech. However, some improvements are observed for lower SNR values when the C_0 coefficient is removed. Note that results obtained for SNR = 20 dB is 2% abolute better that results for clean speech (i.e. SNR = Inf). This can be explained by the higher SNR mean value of the training data compare to the data used for the evaluation. Thus, adding some noise before decoding leads to a better matching between the testing and the training set.

[3] cf. Section 4.4

Models \ SNR	Inf	20 dB	15 dB	10 dB
Baseline	31.53	31.26	36.87	55.00
Baseline_noC_0	32.74	30.65	35.68	51.07
Extended_SS_multi_noC_0	30.79	29.80	30.09	34.04
SphinxDenoise_multi_noC_0	32.55	31.58	33.08	36.73
PNCC_mutli	33.29	33.47	35.46	39.46

Table 2. WER (%) according to the Signal-to-Noise ratio (Δ_T = 0.0 s).

Models \ SNR	Inf	20 dB	15 dB	10 dB
Baseline	-	31.03	37.07	55.23
Baseline_noC_0	-	31.34	35.79	51.67
Extended_SS_multi_noC_0	-	33.03	33.64	37.78
SphinxDenoise_multi_noC_0	-	33.88	34.50	39.67
PNCC_multi	-	34.76	36.43	41.74

Table 3. WER (%) according to the Signal-to-Noise ratio (Δ_T = 4.0 s).

Concerning the noise robust feature sets a clear improvement of results according to SNR is observed compared to the baseline model. Between clean and 10 dB SNR, the WER only increases by approximatively 5 % absolute. The model based on the extended spectral subtraction gives better results than the PNCC or the embedded Sphinx denoise features.

4.3. Robustness with respect to non-speech segments

Previous results show the usefulness of removing the C_0 coefficient for increasing robustness to the input level variation. However, this modification may affect the detection and the recognition of speech surrounded by non-speech segments, especially when no VAD is used. To evaluate this aspect, various SNR and non-speech segment durations are used (as described in section 3.5).

Table 3 gives the results for non-speech segments of 4 second duration (Δ_T = 4.0 s). The main result is the slight influence of non-speech region before or after an utterance for baseline features without C_0 (compared to the results with Δ_T = 0.0 s, cf. Table 2, column 2). Concerning the noise robust feature sets, the extended spectral subtraction provides the best performance.

Figure 2 displays results for different values of Δ_T and for a SNR = 15 dB. Here again, for all non-speech segment durations, the extended spectral subtraction features lead to the best results. For all feature sets, a slight performance degradation is observed when the duration of the non-speech segments increases.

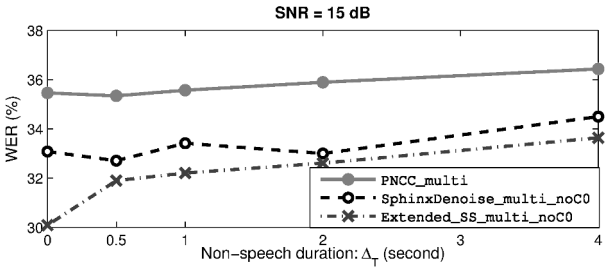

Fig. 2. WER of *PNCC_multi*, *SphinxDenoise_multi_noC*$_0$ and *Extended_SS_multi_noC*$_0$ models for various Δ_T and for SNR = 15 dB.

4.4. PNCC vs. input signal level

Table 1 shows a large performance degradation of the PNCC features when the input signal is attenuated. This is due to the processing of rather shorts speech segments. Decoding short segments prevent the PNCC algorithm to converge to the average power in order to apply a proper power normalization. The initial power parameters (denote as μ_0 in [6]) was initialized to the value returned by the PNCC analysis of the whole signal file from which the segments were extracted. This gives good results for the original modified segments (Att. = 0 dB). But, when the speech signal is attenuated, the initial power value do no match properly. In order to estimate the influence of the initial power parameter on performance robustness with respect to the signal level, the initial power parameter (μ_0) was adjusted according to each attenuation level.

The results of this experiment are presented in Table 1 in the line denoted *PNCC_multi+Adapted_*μ_0. Even if performance is not the best one, adjusting the initial power parameter μ_0 of the PNCC algorithm gives much better results than the PNCC without such adjustment. We could even expect better results with an improved adjustement of μ_0 to each utterance.

Other experiments (not reported here) have shown that at low SNR values, when PNCC features are computed on long duration speech signals, these features provide the best recognition performance. However, the reported experiments show that the setting of this initial power parameter has a critical importance on the speech recognition performance when dealing with rather short speech segments (i.e., the size of an utterance).

5. CONCLUSION

This paper has evaluated the robustness of different approaches for speech recognition with respect to perturbations that are frequent in real operating conditions. The experiments showed that removing the C_0 coefficient, in MFCC-based features, does not impact a lot on the speech recognition performance in clean speech condition nor when non-speech segments are present before or after the actual utterance to recognize, and this makes the speech recognition system much more robust with respect to the input speech signal level. Results also show that the extended spectral subtraction approach, including corrections, gives the best results which are slightly better than the sphinx embedded denoising approach, and slightly better than the baseline, even in clean speech conditions. Results also show that PNCC features are a promising set of noise robust features, but they are strongly dependent on the initialization of the initial power parameter, especially when applied on rather short duration speech segments. Future work will investigate solutions for a better initialization of this parameter, to make the PNCC approach suitable for real time speech recognition, even on short duration speech segments.

REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 4(22), pp. 745–777, 2014.

[2] A. Pawar, K.B. Choudhari, and M.A. Josh, "Review of single channel speech enhancement methods in spectral domain," *Int. J. of Applied Engineering, Research (IJAER)*, vol. 7, no. 11, pp. 1961–1966, 2012.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 1979, pp. 208–211.

[4] H. Hermansky, "Perceptual linear prediction analysis of speech," *J. Acoust. Soc. Am.*, vol. 87(4), pp. 1738–1752, 1990.

[5] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 2(4), pp. 578–589, 1994.

[6] C. Kim, *Signal processing for robust speech recognition motivated by auditory processing*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA USA, 2010.

[7] P. Sovka, P. Pollak, and J. Kybic, "Extended spectral subtraction," in *Proc. of the 5th European Conference on Speech Communication and Technology*, 1995, pp. 963–966.

[8] B. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] S. Galliano, G. Gravier, and L. Chaubard, "The Ester2 evaluation campaign for rich transcription of french broadcasts," in *Proc. of Interspeech*, 2009.

[10] G. Gravier, G. Adda, N. Paulson, M. Carr, A. Giraudel, and O. Galibert, "The Etape corpus for the evaluation of speech-based TV content processing in the french language," in *Language Resources and Evaluation (LREC'12)*, 2012.

[11] Y. Estve, T. Bazillon, J.-Y. Antoine, F. Bchet, and J. Farinas, "The Epac corpus: Manual and automatic annotations of conversational speech in french broadcast news," in *Language Resources and Evaluation (LREC'10)*, 2010.

[12] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 hub-4 sphinx-3 system," in *Proc. of the 1997 ARPA Speech Recognition Workshop*, 1997, pp. 85–89.

[13] E. Principi, S. Squartini, F. Piazza, D. Fuselli, and M. Bonifazi, "A distributed system for recognizing home automation commands anddistress calls in the italian language," in *Proc. of Interspeech*, 2013, pp. 2049–2053.