Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise

Thomas Schatz[1,2]*, Vijayaditya Peddinti*[3]*, Xuan-Nga Cao*[1]*, Francis Bach*[2]*,*
Hynek Hermansky[3]*, Emmanuel Dupoux*[1]

[1] LSCP, ENS/EHESS/CNRS, Paris, France
[2] SIERRA Project-Team, INRIA/ENS/CNRS, Paris, France
[3] HLT Center of Excellence, John Hopkins University, Baltimore, Maryland

thomas.schatz@ens.fr, vijay.p@jhu.edu, ngafrance@gmail.com, francis.bach@ens.fr,
hynek@jhu.edu, emmanuel.dupoux@gmail.com

Abstract

The Minimal-Pair ABX (MP-ABX) paradigm has been proposed as a method for evaluating speech features for zero-resource/unsupervised speech technologies. We apply it in a phoneme discrimination task on the Articulation Index corpus to evaluate the resistance to noise of various speech features. In Experiment 1, we evaluate the robustness to additive noise at different signal-to-noise ratios, using car and babble noise from the Aurora-4 database and white noise. In Experiment 2, we examine the robustness to different kinds of convolutional noise. In both experiments we consider two classes of techniques to induce noise resistance: smoothing of the time-frequency representation and short-term adaptation in the time-domain. We consider smoothing along the spectral axis (as in PLP) and along the time axis (as in FDLP). For short-term adaptation in the time-domain, we compare the use of a static compressive non-linearity followed by RASTA filtering to an adaptive compression scheme.

Index Terms: noise resistance, zero-resource, speech features, evaluation framework, minimal-pair ABX task

1. Introduction

Speech features are typically evaluated through their effect on an entire speech recognition system through phoneme-error-rates or word-error-rates. These metrics are problematic in zero-resource/unsupervised settings (were only very limited amount of labeled data, if any, is available for training [1, 2]). First, they are expensive, since a speech recognition system needs large amounts of hand-labeled speech data to be trained. Second, they lack sensibility in that the supervised training of the recognition system might compensate for potential defects of the speech features (such as noisy or unreliable channels for example), even though such defects can be very harmful to zero-resource applications. Finally, they are hard to interpret, since typical speech recognition pipelines are complex models with potential biases in favor of specific types of features that are hard to understand and predict (a simple example is that of diagonal-covariance GMM-HMM models, which are biased toward de-correlated features).

In [3], we proposed an alternative to phoneme-error-rates for evaluating speech features: the error rate in a minimal-pair ABX task (MP-ABX task). A MP-ABX task exploits the simple idea that in order to understand a language, it is necessary to *discriminate* between minimal-pairs of words from this language. In an MP-ABX task, the features a, b and x associated to three speech sounds, A, B and X are computed, where A and B are chosen to be minimally different words (e.g. dog vs doll) and X is linguistically identical to either A or B, although it can be indexically different (different talker or added noise for example). Then, one determines whether x is closer to a or b according to a metric defined on the space of the evaluated features, and the result is compared to the expected answer. By repeating this on a representative set of A, B, X triplets, a measure of the discriminability of minimal pairs when coded with the tested featural representation is obtained. This evaluation metric is especially suitable for zero-resource settings as it doesn't unduly correct defects in the speech features and it encapsulates all modeling assumptions in the choice of a metric on the space of the features, an object conceptually much simpler than a typical speech recognition pipeline.

In [3], all MP-ABX evaluations were performed with clean speech. Noisy conditions, however, represent a major challenge for speech recognition technologies. The introduction of techniques based on deep neural networks [4, 5], recently have lead to some breakthroughs in performance, questioning the usefulness of more traditional approaches to noise-robust speech features extraction (e.g. [6]). However, these techniques require large amounts of hand-annotated speech signal and thus aren't appropriate in zero-resource settings. In this study, we consider signal processing operations commonly used for the extraction of noise-robust speech features and analyze their performance in MP-ABX tasks with noisy stimuli. More specifically, we perform an experiment with additive noise and another with convolutional noise. In Experiment 1, car and babble noise from the Aurora-4 database as well as white noise are added to Consonant-Vowel (CV) stimuli from the Articulation Index corpus [7] at various signal-to-noise ratio (SNR). For each type of noise and at each SNR, we evaluate the features in a cross-speaker MP-ABX task. In Experiment 2, we create stimuli with an emphasis on the high frequencies or on the low frequencies by convolution of the CV stimuli with a high-pass and a low-pass filter respectively. The features are evaluated in two different tasks for this experiment: a cross-speaker MP-ABX task *within* convolutional noise, to measure the ability of features to *adapt* to a particular distortion, and a cross-speaker MP-ABX task *across* convolutional noise; to measure the ability of features to *be invariant* to convolutional distortions.

We evaluate two class of noise-robustness techniques: techniques that work by smoothing a time-frequency representation

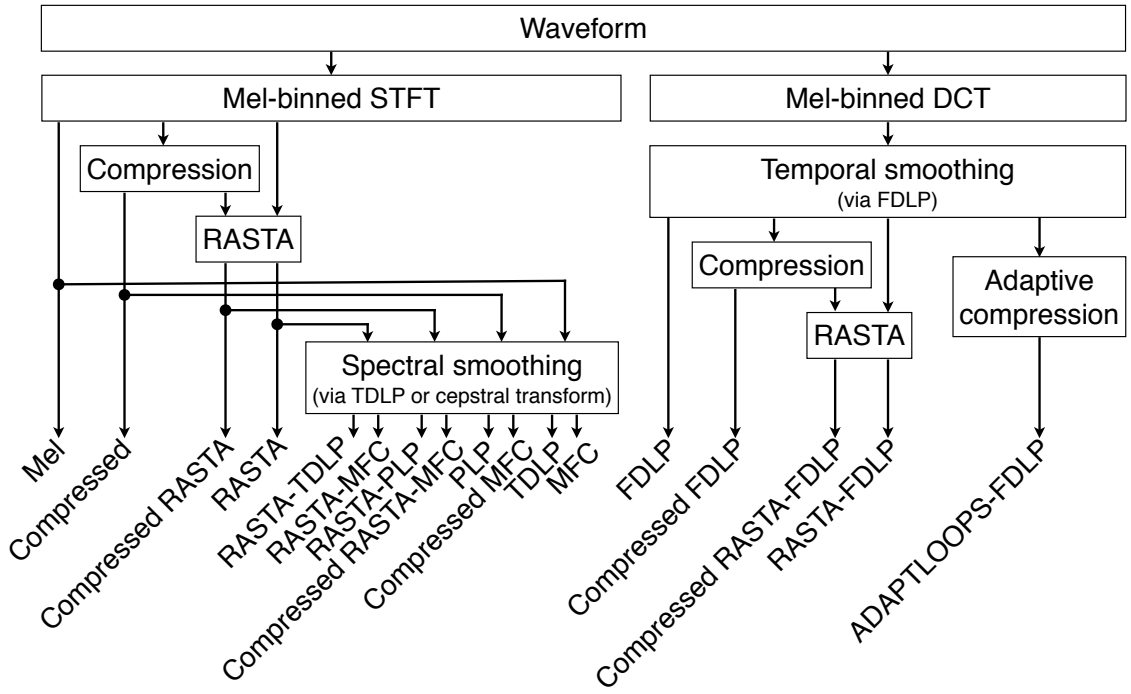

Figure 1: *The 17 speech feature extraction pipelines tested in this paper.*

along the frequency axis (as in PLP [8] or MFC [9]) or the time axis (as in FDLP [10]) and techniques that enhance transients in the time-domain and/or compress the dynamic-range in each spectral channels [12, 13, 14].

2. Methods

2.1. Stimuli

We use a subset of the Articulation Index Corpus (LDC-2005S22) [7], consisting of all possible CV syllables of American English pronounced inside a carrier sentence by 3 male and 3 female speakers for a total of 1709 stimuli recorded with a sampling rate of 16KHz and manually time-aligned by Xuan-Nga Cao. For additive noise, we draw a random sample from the car and babble noise of the Aurora-4 database or we generate a white noise sample of the length of the carrier sentence, and we scale it to obtain the desired SNR by using only the parts of both the signal and the noise corresponding to the carried CV. For convolutional noise, we use bidirectional filtering with two different order one Butterworth filters: a low-pass filter with a cutoff frequency of 100Hz and a high-pass filter with a cutoff frequency of 4000Hz.

2.2. MP-ABX Tasks

A phoneme discrimination MP-ABX task [3] is used, where A and B differ in only one phoneme and X is identical to either A or B. The task is done across talkers (A and B are uttered by the same talker and X by a different one) and within noise (the same kind of noise is applied to A, B, and X and, for additive noise, at the same SNR). In experiment 2, another MP-ABX task is also used. This second task is identical to the first in every respect, excepted that it is not only *across speaker* but also *across noise*, in the sense that either A and B are noisy and X is noiseless or A and B are noiseless and X is noisy.

2.3. Data Analysis

For each triplet A, B, X, the sign of $d(a, x) - d(b, x)$, is used to determine the response of the model (B if positive, else A), where a, b, x are speech features for A, B, X and d is a Dynamic-Time-Warping distance based on a frame-level cosine distance [11, 3]. For each task, the average error over all the possible triplets of stimuli A, B and X respecting the constraints of the task is computed.

2.4. Speech features

All the representations we use are Mel-scale time-frequency representations sampled every 10ms along the time axis with 21 spectral channels. We compare a basic Mel-scale power spectrum to a Mel-scale spectrum smoothed using either linear predictive coding in the time-domain (TDLP) or in the spectral domain (FDLP), or a cepstral transform (MFC), or linear predictive coding in the time domain after having applied a cubic-root compression of the dynamic range in each spectral channel (PLP). We don't perform equal-loudness filtering when computing PLP features as we previously found that it was harming the performance in the MP-ABX task [3]. We tested several values for the order of the linear predictive coding models for TDLP, PLP and FDLP and for the number of cepstral coefficients used for reconstructing the spectrum for MFC (cf. Table 1). We also consider representations derived from a Mel-scale spectrum using cubic-root compression of the dynamic range (Compressed) or using RASTA-filtering (RASTA) [12] or both (Compressed RASTA) and all relevant combinations of these with FDLP, TDLP or MFC smoothing. We also tested a representation obtained using an adaptive compression scheme (ADAPTLOOPS-FDLP) that performs both dynamic-range compression and short-term adaptation in the temporal domain in one step [13, 14].

3. Results

In [3], we used only CV recorded in isolation, here we use sentence-embedded CV instead. Indeed, the performance on isolated stimuli can be misleading about the performance of some of the techniques on continuous speech. This is illustrated in the case of RASTA filtering in Figure 2, where we compare a RASTA-filtered Mel-spectrum to a plain Mel-spectrum in the cross-talker MP-ABX task. The large benefit of RASTA filtering with the isolated stimuli disappear almost entirely with the sentence-embedded stimuli.

For each of the signal processing pipelines depending on a parameter (model order or number of cepstral coefficients), the value of the parameter that yielded the best performance on average over all additive noise types and all SNR was selected (cf. Table 1). All results reported in the following are obtained using these parameter values.

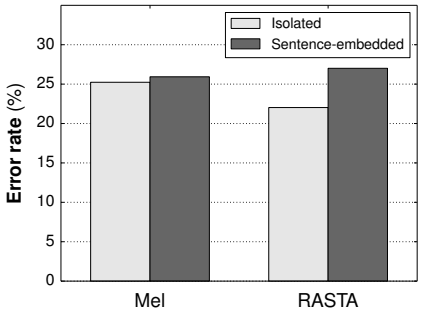

Figure 2: *MP-ABX error rate for the Mel baseline and the RASTA pipeline, tested with syllables in isolation versus extracted from a carrier sentence.*

3.1. Additive Noise

The error rate in the cross-speaker MP-ABX, averaged over the three types of additive noise, is plotted for various speech features in Figure 3 a, b, c as a function of SNR. Features operating by smoothing on a time-frequency representation are compared to a simple Mel-spectrum in Figure 3 a. No feature appears best for all SNR and, while TDLP is not improving very much on the Mel-spectrum baseline, both MFCC and PLP are much better for high SNR and PLP remain much better for low SNR as well. FDLP is worse than a simple Mel-spectrum in clean conditions, but it's by far the best feature for low SNR. In Figure 3 b, features operating by enhancement of time-domain transients and compression of the dynamic range are compared. The adaptive compression scheme has very poor performance for high SNR conditions and an average performance for low SNR. A simple compressed Mel-spectrum gives very good performance for high SNR, but becomes worse than a RASTA-filtered Mel-spectrum for low SNR. The compressed RASTA-filtered Mel-spectrum is a good compromise between the two. In Figure 3 c, features with the best performance on average for all additive noise (cf. Table 2) are compared. Interestingly, no single tech-

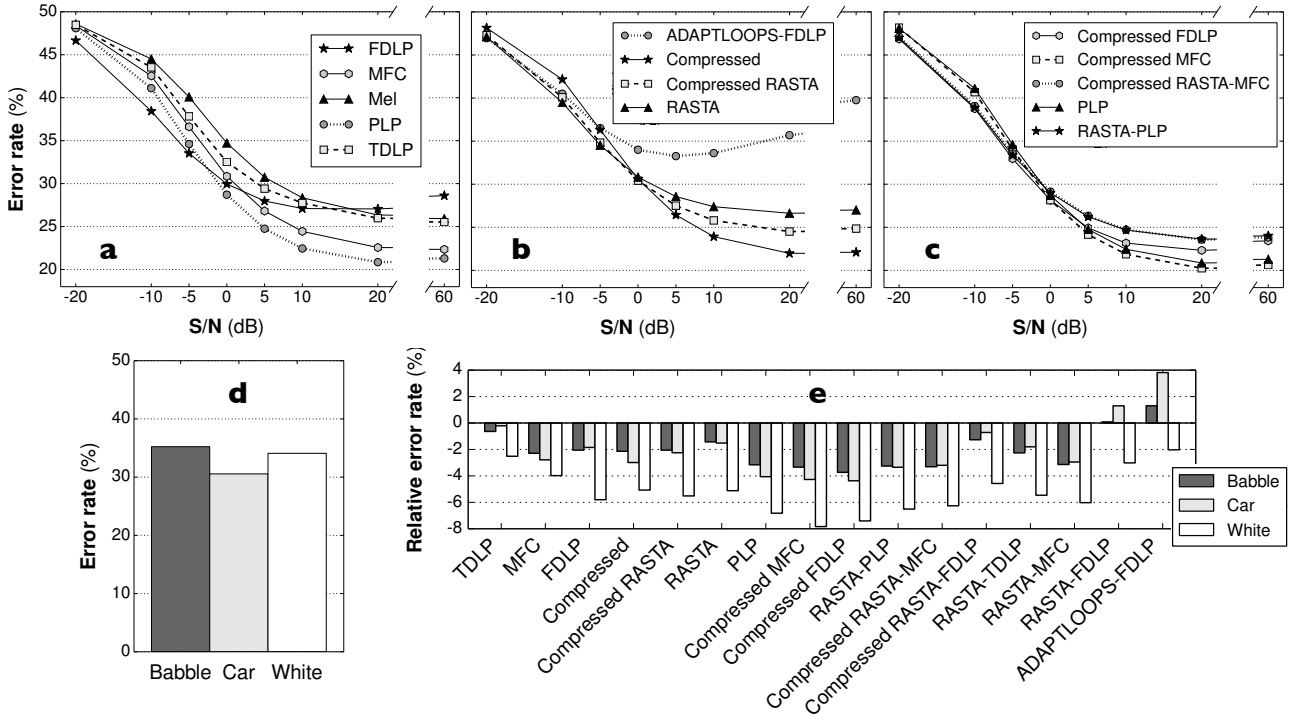

Figure 3: *a, b and c: MP-ABX error rates for various speech features as a function of Signal to Noise ratio (in dB). Clean speech corresponds to 60 dB, and the chance level is 50%. a: MP-ABX error rates for the Mel baseline and for techniques performing a form of smoothing of the time-frequency representation. b: MP-ABX error rates for techniques involving short-term adaptation and/or dynamic-range compression in the frequency channels. c: MP-ABX error rates for the five techniques that perform best in additive noise, according to Table 2. d: average MP-ABX error rate over all SNR for each additive noise for the Mel baseline. e: difference in MP-ABX error rate between the Mel baseline and the sixteen other pipelines. Negative values indicates that the pipeline performs better than baseline.*

Pipelines	Parameter							
FDLP	2	3	5	8	**13***	20	32	50
TDLP	2	**3***	5	7	12	20		
MFC	2	3	**5***	7	12	20		
PLP	2	3	5	**7***	12	20		
RASTA-FDLP	2	3	5	**8***	13	20	32	50
RASTA-TDLP	2	**3***	5	7	12	20		
RASTA-MFC	2	3	5	**7***	12	20		
RASTA-PLP	2	3	5	**7***	12	20		
Compressed FDLP	2	3	5	8	13	**20***	32	50
Compressed MFC	2	3	**5***	7	12	20		
Compressed RASTA-FDLP	2	3	5	8	**13***	20	32	50
Compressed RASTA-MFC	2	3	5	**7***	12	20		
ADAPTLOOPS-FDLP	2	3	5	**8***	13	20	32	50

Table 1: *Values tested for the order of linear predictive coding models/the number of cepstral coefficients of features. For FDLP-based features the order corresponds to the number of poles per second of signal. The parameters in bold with an asterisk yielded the best overall performance on the cross-talker MP-ABX task.*

nique appears to dominate all others at both low and high SNR at the same time.

The differences between the various additive noises are highlighted in Figure 3 d and e. For a simple Mel-spectrum representation, car noise has the least impact on the discriminability of phonemes, white noise has the most impact and babble noise is intermediate as shown in Figure 3 d . Interestingly, Figure 3 e shows that although white noise has the biggest impact on the discriminability of phonemes based on a Mel-spectrum, it is also the most easily compensated for. In contrast babble noise, that is already quite damaging on a Mel-spectrum is much harder to compensate for. Notice also that most of the time features that work well for compensating a particular type of additive noise, also work well for compensating the others.

3.2. Convolutional Noise

The results regarding convolutional noise are shown in Table 2. The error rates for the *within noise, across-talker* MP-ABX task are in the *Within* column and those for the *across noise, across-talker* MP-ABX task are in the *Across* column. Overall, pipelines incorporating smoothing or short-term adaptation along the time-axis (i.e. incorporating FDLP or RASTA respectively) appear more performant for the *within noise, across-talker* MP-ABX task and vastly more performant for the *across noise, across-talker* task. Interestingly the gains of using FDLP and RASTA do not add up when combining the two. This may be explained by the opposite modes of action of FDLP and RASTA filtering: FDLP extracts a slowly-varying envelope in each spectral channel while RASTA-filtering remove the slowly varying components from these channels.

The most performant pipeline on convolutional noise is the Compressed-FDLP pipeline for both tasks. It is also the most performant pipeline overall for additive noise, along with the Compressed MFC pipeline. Its only weak point is for clean speech coding, where several other pipelines fare better.

4. Conclusion

We built upon previous work by Schatz et al. [3] to propose a new framework for the evaluation of the resistance to noise of speech representations in the zero-resource/unsupervised settings. We used cross-talker MP-ABX tasks to probe 17 pipelines, that implemented different schemes for noise-resistance. We found, as we expected, that pipelines that incorporated spectral or temporal smoothing of a time-frequency representation were among the most resistant to additive noise, while pipelines that incorporated short-term adaptation or temporal smoothing were among the most resistant to convolutional noise. The Compressed FDLP features were optimal for both kind of degradations. However, no single pipeline was optimal for additive noise at all SNR: some pipelines were more resistant than others to extreme signal degradation, but they were also less performant in the absence of degradation. Further work is needed to explore whether a combination of the techniques we investigated in this paper or of other common techniques for robust feature extraction can yield features that would be optimal in all situations, or whether techniques based on actively estimating the properties of the environment and adapting the features to it would fare better in zero-resource/unsupervised applications.

Table 2: *Error rates for the 17 pipelines: in the across-talker MP-ABX task for additive noise (grand average across the three noise types and seven S/N ratios); in the across-talker, **within noise** MP-ABX task for convolutional noise and in the across-talker, **across noise** MP-ABX task for convolutional noise (averaged over high frequency and low frequency emphasis); in the across-talker MP-ABX task for clean speech.*

Pipelines	Noise			Clean
	Additive	Convolutional		
		Within	Across	
Baseline				
Mel	36.2	41.7	43.0	25.9
Smoothing				
FDLP	33.0	28.4	29.0	28.6
TDLP	35.1	38.0	45.9	25.5
MFC	33.2	35.7	43.5	22.3
Compression/Adaptation				
RASTA	33.5	27.6	28.8	27.0
Compressed	32.8	32.2	41.8	22.1
Compressed RASTA	32.9	25.2	26.7	24.8
Smoothing + Compression/Adaptation				
PLP	31.5	31.3	42.3	21.3
RASTA-PLP	31.8	24.4	25.9	24.0
RASTA-MFC	32.2	25.9	27.5	25.3
RASTA-TDLP	33.0	27.6	28.9	26.9
RASTA-FDLP	35.6	32.1	32.4	32.2
Compressed MFC	**31.0***	28.3	42.7	**20.6***
Compressed FDLP	**31.0***	**23.5***	**24.2***	23.4
Compressed RASTA-MFC	31.9	24.1	26.0	23.8
Compressed RASTA-FDLP	34.0	28.4	28.7	28.5
ADAPTLOOPS-FDLP	37.2	38.8	39.5	39.8

5. Acknowledgements

The research leading to these results received funding from the European Research Council under the FP/2007-2013 program / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-0001-02 PSL*, ANR-10-LABX-0087) and from the Fondation de France.

6. References

[1] J. Glass, "Towards unsupervised speech processing," in *2012*, 2012, pp. 1–4.

[2] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proceedings of ICASSP 2013*, 2013, pp. 8111–8115.

[3] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hynek, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *INTERSPEECH-2013*. International Speech Communication Association, 2013, pp. 1781–1785.

[4] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009. [Online]. Available: http://www.cs.utoronto.ca/ gdahl/papers/dbnPhoneRec.pdf

[5] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, p. 73987402.

[6] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," *Techniques for Noise Robustness in Automatic Speech Recognition*, p. 193227, 2012.

[7] P. Fousek, P. Svojanovsky, F. Grezl, and H. Hermansky, "New nonsense syllables database analyses and preliminary asr experiments," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2004, pp. 2004–29.

[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[9] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 91–103, 1976.

[10] S. Ganapathy, S. Thomas, and H. Hermansky, "Static and dynamic modulation spectrum for speech recognition,," in *Proceedings of Interspeech*, 2009.

[11] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proceedings of Interspeech*, 2011.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[13] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[14] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.