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$^4$, 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 robustness to additive 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 sensitivity 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], 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 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 convolutional noise. In Experiment 2, we evaluate two class of noise-robustness techniques: techniques that work by smoothing a time-frequency representation...
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.

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-
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 would be optimal in all situations, or whether techniques for robust feature extraction can yield features that were also less performant in the absence of degradation.

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-resistant. 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.

## 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


