How Realistic is Artificially Added Noise?

Thomas Winkler
Fraunhofer IAIS, Sankt Augustin, Germany
thomas.winkler@iais.fraunhofer.de

Abstract
Evaluations of algorithms for robust automatic speech recognition (ASR) are often based on artificial noisy speech instead of realistic noisy speech. In this paper we compare the ASR performance of speech with artificial additive noise to the performance of realistic noisy speech. All data was recorded during the same recording campaign and with nearly identical channel characteristics. The simulation process takes into account all major characteristics of the noisy reference data. Clean speech, noisy speech and simulated speech are compared for different aspects of robust ASR including noise reduction by Spectral Subtraction and the ETSI robust front end. The results show, that artificial noisy speech even in very controlled simulation environments is not very similar and not a full substitute for realistic noisy data. While the tendencies of the improvement for artificial and realistic data are similar for the evaluated approaches, the magnitude can be quite different.

Index Terms: additive noise, simulated noise, robust ASR

1. Introduction
Artificial noisy speech data helps us evaluating and benchmarking different aspects and algorithms of robust automatic speech recognition. We often assume that noise artificially added to clean speech properly represents a realistic noisy environment. While some assumptions made for creating such data are approximations, other influences on realistic noisy data are often not even considered during the simulation process. Thus, the performance of certain algorithms analysing a signal recorded in realistic noise conditions might be slightly or even completely different compared to the performance of the same algorithms on a signal with artificially added noise. We can assume that this mainly depends on the quality of the simulation and the nature of the algorithm — especially if an algorithm is based on an assumption which is true for the artificial noisy data but not for realistic data or vice versa.

Certain advantages of simulated noise are obvious: one can easily evaluate the effect of background noise added to a clean speech signal as both sources — clean speech and background noise sample — are known and the degree of mixing can be controlled precisely. In other cases a sufficient amount of noisy speech with certain background noise might be required but not available. Here artificial noisy speech is usually much easier to obtain than realistic noisy speech, which often has to be recorded and prepared expensively. A prominent example for an evaluation database based on artificial noisy data is AURORA-2 [1], where typical background noise of different locations is added to clean speech samples of the TIDIGITS database. Beyond additive noise channel characteristics and other effects can also be modelled and simulated, e.g., coding algorithms and filter functions common in mobile phone communications [2].

In most simulations of realistic noisy speech, clean speech and environmental noise are assumed to be independent. But especially for certain types of noise and low signal-to-noise ratios (SNRs) this assumption is not necessarily valid, as the Lombard effect, for example, influences the way the speaker is speaking dependent on the level and type of environmental noise [3, 4]. Furthermore, the different channel characteristics and the room response of the source recordings are often neglected when adding noise samples recorded in one environment with a certain hardware setup to speech recordings from another environment recorded with a different hardware setup. Whenever an algorithm is based on or dependent on any of these assumptions, artificial noise might not be as realistic as desired and results of an evaluation based on this data loose their significance. Still, artificial noise is very useful to understand and evaluate certain aspects in a laboratory environment. So we know that artificial noisy speech is important in scientific research, but we wonder, how realistic artificial noise finally is. In this paper we try to answer this question for a standard approaches of automatic speech recognition (ASR) with and without robust feature extraction.

In Section 2 we introduce the basic mathematical model for additive noise. Section 3 describes the simulation process and dataset, while Section 4 describes the evaluation setup. The results are presented, discussed, and concluded in Section 5 and Section 6.

2. Additive Noise
Both, algorithms for noise simulation and approaches for noise reduction, often assume a linear combination of speech and background noise. Noise reduction by Spectral Subtraction [5, 6], for example, is one of the most prominent groups of algorithms strongly based on the assumption of purely additive noise. Assuming that noise and speech are independent, noisy speech $\tilde{s}(t)$ can be described by adding a noise signal $n(t)$ to a clean speech signal $s(t)$. But in realistic data usually nonlinear effects of the room response or the channel characteristics can significantly influence the final signal. Thus such effects should also be considered to improve the mathematical model of recorded noisy speech. These effects can be described by a convolution in temporal domain with the transfer function $h(t)$ (with the convolution operator $*$):

$$\tilde{s}(t) = [s(t) + n(t)] * h(t) = \tilde{s}(t) + \tilde{n}(t) \quad (1)$$

with

$$\tilde{s}(t) = s(t) * h(t) \quad (2)$$

$$\tilde{n}(t) = n(t) * h(t). \quad (3)$$

In the special case of the same transfer function for all recorded signals, i.e., same room response and channel characteristics for clean speech, noisy speech, and noise samples, we can use the right hand side of Equation 1 to simply add our recorded noise.
samples \( \tilde{s}(t) \) to the recorded clean speech \( \hat{s}(t) \). This only applies for signals recorded in identical environments and setups where the transfer function \( h(t) \) is time invariant and identical for all signals. For any arbitrary noise and speech samples this is usually not the case. But for the datasets used in this work we can consider this assumption to be valid as we will discuss in the following section.

3. Evaluation Data

We use the MoveOn Motorcycle Speech Corpus described in [7] which contains clean speech, realistic noisy speech and realistic noise samples. The noisy speech and the pure noise samples were recorded on the motorcycle while driving. The additional clean speech utterances were recorded in an office environment but using the same helmet and hardware setup. Speech is orthographically transcribed and background noise samples are annotated with occurring noise types. The most dominant noise types (engine and air wind noise) were annotated in three levels of magnitude (low, medium, high).

In the MoveOn Motorcycle Speech Corpus the same hardware and recording setup for all recorded signals were used. The only difference between clean speech and noisy speech recordings were the different speakers as well as the office environment as opposed to the outdoor environment. In case of this database the effect of the different environments — in terms of the room response but not the background noise — can be considered minimal, as the direct environment was always the full face motorcycle helmet that all speakers had to wear during both office and motorcycle recording sessions. Thus, channel characteristics and room response are practically identical for all recordings. In case of the different speakers only native speakers of British English were selected. Furthermore, evaluation is performed strictly speaker independent to avoid any effect of speaker dependency appearing in one but not the other evaluation set.

3.1. Data Characteristics

Each reference noisy speech segment is described by its duration, speaker, transcription, type of background noise, and estimated SNR. Clean speech segments are also referred to by duration, speaker, and transcription but without any noise related information. All noise samples contain nearly homogeneous noise conditions for stationary noise types (traffic, engine noise of the same level, etc.) respectively exactly encapsulate one noise event for non-stationary noises (passing cars, horns, etc.). The noise samples are characterised by duration and type of background noise. All available metadata is considered for the simulation of the noisy speech data as described in Section 3.3, except for the speaker information which we only use to build training and test sets for a strictly speaker independent evaluation.

3.2. SNR Estimation

We estimate the SNR of the reference data using NIST segsnr from the NIST Quality Assurance Package\footnote{NIST SPQA 2.3, http://www.itl.nist.gov/iad/mig/tools/} in combination with a voice activity detection (VAD) based on the alignment algorithm of HTK\footnote{HTK - Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/}. As we have a full orthographic transcription of all speech segments, training of well adapted acoustic models and automatic alignment with HTK become possible. The training process and the performance of the acoustic models is described in [8]. Each utterance is aligned on a phoneme basis using HTK with the acoustic models and transcriptions. All silence and non-speech monophones are clustered as non-speech, while all other phonemes are clustered as speech. The time information from the alignment is used to provide the voice activity information for the SNR estimation. We expect an improved VAD compared to blind algorithms by using this additional knowledge from the transcriptions and the acoustic modelling. The voice activity information is finally used with NIST segsnr to estimate the SNR for each noisy utterance.

3.3. Simulation of Noisy Speech

Artificial noisy speech is generated based on the statistics and information of the noisy reference data. We use two simulation passes to be able to adapt the SNR between the first and second iteration. This is necessary, as the resulting simulated and the reference noisy utterances can only be compared by their estimated SNRs and not by the precise SNR, which could only be calculated for the artificial data. We followed the following process for simulating realistic noisy speech:

1. Selecting speech sample. For each reference utterance a clean speech utterance with a similar phoneme sequence is selected. The similarity of the phoneme sequence is measured by the accuracy of a clean speech phoneme sequence matching the reference phoneme sequence. We select one of the previously unused clean speech utterances with the highest accuracy.

2. Selecting noise sample. For each reference utterance a noise sample is selected that fulfills two requirements. First, the noise annotation of a noise sample must be identical to the noise annotation of the reference utterance; second, the noise sample must be at least as long as the clean speech utterance selected in Step 1.

3. Mixing samples. The selected clean speech utterance is mixed with a selected noise sample using Equation 1, right hand side. The signal levels of the noise and the speech samples are adjusted before mixing to fit the estimated SNR of the reference noisy speech utterance.

4. Estimating SNR. The SNR for the simulated noisy speech utterance is estimated as described in Section 3.2. When we compare the estimated SNR of the simulated noisy utterance to the estimated SNR of the reference noisy utterance, we can see that it deviates from the ideal line (Figure 1, dotted line vs. straight line).

5. Correcting SNR. We estimate a piecewise approximation of the relation between used (reference) SNR \( SNR_{used} \) and the estimated SNR of the simulated noisy speech \( SNR_{est} \) by polynomial functions up to the order of three (Equation 4) with the parameters found in Table 1. Towards 0dB the estimated SNRs begins to converge due to imprecise estimations at low SNRs. To avoid distortions caused by very low simulated SNRs during the mixing process, the SNR is constantly set to 0dB for all aimed SNRs below 0dB.

6. Re-mixing samples. For each reference utterance we calculate the new SNR value for mixing \( SNR_{used} \) to get a desired estimated SNR \( SNR_{est} \) by using the polynomial approximation in Equation 4 with parameterisation from Table 1. This SNR value is used in a second pass to remix clean speech utterances and noise samples. Figure 1 shows the improvements of the correction.
Each set. Each of the speakers builds the test set in one round, the database, we use cross validation for the six main speakers of the set. The artificial noisy speech sets are created for each training and test set separately as described in Section 3.3. The six major speakers of the clean speech samples are also iterated, so that we have a clean speech test speaker as counterpart for each noisy speech test speaker. Each training set contains about 1400 utterances, while each test set contains about 170 utterances. Each full set consisting of the noisy reference data with its corresponding simulated and clean speech data is referred to as cross validation set, while clean, simulated and noisy speech are called attributes in the following text.

We evaluate a standard ASR system based on Mel Frequency Cepstral Coefficients (MFCCs). We use 39 features including the first 12 MFCCs plus energy and first and second order derivatives. We also evaluate two noise reduction and robust feature extraction algorithms, namely an approach based on Spectral Subtraction [5] as well as the robust front end standardised by ETSI [9]. The first approach uses Spectral Subtraction with noise estimation introduced in [10] and Wiener filtering. We chose this as a representative approach based on the assumption we used to artificially mix clean speech and noise samples (Section 2). We used the algorithm specsub.m v 1.4 for Matlab provided in voicebox with default parameters but using Wiener filtering. The ETSI robust front end represents a more complex feature extraction front end for mismatch reduction. It also contains a Wiener filter based noise reduction but with additional mismatch compensation steps. The ETSI robust front end represents a quite powerful standardised approach for robust feature extraction.

All acoustic models are trained using HTK. For training and testing the plain approach, features were extracted from plain audio with HTK. For the Spectral Subtraction evaluation Spectral Subtraction was performed on the audio before extracting features with HTK. The ETSI features are directly provided by the ETSI robust front end. For all settings we consider a set of 44 phonemes (plus silence and short pause model) for British English. For each phoneme a Hidden Markov Model with 16 Gaussian mixtures per state is trained. Test and training data are disjoint. All combinations of training and test data are strictly speaker independent.

### 5. Results

All three approaches in the previous section are evaluated measuring the phoneme accuracy of an ASR system using only the pure acoustic information without incorporating any lexical knowledge. Each attribute (clean, simulated, noisy) is evaluated on all acoustic models of the same cross validation set. The means of the phoneme accuracy of all cross validation sets are shown in Table 2.

For a plain recognition all acoustic models perform best on the test set of the same attribute Table 2, mfcc). The simulated data shows a comparatively low performance evaluated on its own acoustic models. On the other hand, the acoustic models based on simulated data show the best performance on test data with different attributes (clean, noisy). This can be explained by the ambivalent character of this set which was acquired by mixing clean speech with background noise. Furthermore, simulated and noisy test sets perform neither equally nor almost equally good on the acoustic models of the respectively other at-

### Table 1: Parameters for polynomial approximation in Equation 4.

<table>
<thead>
<tr>
<th>SNR_est/dB</th>
<th>a3</th>
<th>a2</th>
<th>a1</th>
<th>a0</th>
</tr>
</thead>
<tbody>
<tr>
<td>... &lt; 6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6 ≤ ... ≤ 10</td>
<td>0.0221</td>
<td>-0.8716</td>
<td>12.269</td>
<td>-44.742</td>
</tr>
<tr>
<td>... &gt; 10</td>
<td>0</td>
<td>0</td>
<td>1.1733</td>
<td>1.1071</td>
</tr>
</tbody>
</table>

![Figure 1: Correction of SNR. The resulting estimated SNR for the simulated data and the aimed SNR are plotted. In a first step the reference SNR (aimed SNR) of the noisy data is used directly (●), in a second step the corrected value is used (△). The straight line represents an ideal mapping.](image-url)

$$SNR_{est/dB} = a_3 \frac{SNR_{est}^3}{dB^3} + a_2 \frac{SNR_{est}^2}{dB^2} + a_1 \frac{SNR_{est}}{dB} + a_0$$  \hfill (4)

This process provides a simulated noisy speech dataset with very similar characteristics to the realistic noisy speech dataset. Note that the exact SNR is not relevant here as we are not interested in the SNR value but in similar data sets.

In the following we use the attributes clean, simulated and noisy for specifying entities related to clean speech data, artificially noisy speech and noisy speech data respectively.

### 4. Experimental Setup

If noisy data and artificially noisy data are similar enough, acoustic models trained on one set should perform almost equally well on the test sets of both origins. Furthermore, noise reduction and robust feature extraction algorithms should improve the results testing simulated and noisy data on clean speech models by a similar magnitude.

For evaluation we use the right microphone channel of the MoveOn database. The noisy speech utterances — speech utterances with at least one annotated background noise type — are split into two data sets, each set containing different speakers. As the number of utterances in each set is rather low (limited by the number of available clean speech utterances in the database), we use cross validation for the six main speakers of each set. Each of the speakers builds the test set in one round, while all other speakers of the set are used for training the acoustic models in the same round. Thus, we get two times six reference training and test sets each with disjoint sessions and speakers. The artificial noisy speech sets are created for each training and test set separately as described in Section 3.3. The six major speakers of the clean speech samples are also iterated, so that we have a clean speech test speaker as counterpart for each noisy speech test speaker. Each training set contains about 1400 utterances, while each test set contains about 170 utterances. Each full set consisting of the noisy reference data with its corresponding simulated and clean speech data is referred to as cross validation set, while clean, simulated and noisy speech are called attributes in the following text.

We evaluate a standard ASR system based on Mel Frequency Cepstral Coefficients (MFCCs). We use 39 features including the first 12 MFCCs plus energy and first and second order derivatives. We also evaluate two noise reduction and robust feature extraction algorithms, namely an approach based on Spectral Subtraction [5] as well as the robust front end standardised by ETSI [9]. The first approach uses Spectral Subtraction with noise estimation introduced in [10] and Wiener filtering. We chose this as a representative approach based on the assumption we used to artificially mix clean speech and noise samples (Section 2). We used the algorithm specsub.m v 1.4 for Matlab provided in voicebox with default parameters but using Wiener filtering. The ETSI robust front end represents a more complex feature extraction front end for mismatch reduction. It also contains a Wiener filter based noise reduction but with additional mismatch compensation steps. The ETSI robust front end represents a quite powerful standardised approach for robust feature extraction.

All acoustic models are trained using HTK. For training and testing the plain approach, features were extracted from plain audio with HTK. For the Spectral Subtraction evaluation Spectral Subtraction was performed on the audio before extracting features with HTK. The ETSI features are directly provided by the ETSI robust front end. For all settings we consider a set of 44 phonemes (plus silence and short pause model) for British English. For each phoneme a Hidden Markov Model with 16 Gaussian mixtures per state is trained. Test and training data are disjoint. All combinations of training and test data are strictly speaker independent.

### 5. Results

All three approaches in the previous section are evaluated measuring the phoneme accuracy of an ASR system using only the pure acoustic information without incorporating any lexical knowledge. Each attribute (clean, simulated, noisy) is evaluated on all acoustic models of the same cross validation set. The means of the phoneme accuracy of all cross validation sets are shown in Table 2.

For a plain recognition all acoustic models perform best on the test set of the same attribute (Table 2, mfcc). The simulated data shows a comparatively low performance evaluated on its own acoustic models. On the other hand, the acoustic models based on simulated data show the best performance on test data with different attributes (clean, noisy). This can be explained by the ambivalent character of this set which was acquired by mixing clean speech with background noise. Furthermore, simulated and noisy test sets perform neither equally nor almost equally good on the acoustic models of the respectively other at-

3http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html


6. Conclusion

Many approaches are evaluated on simulated noisy data assuming that the results are valid for realistic data as well. In this paper we created a set of artificial noisy utterances with very similar characteristics concerning speech and noise background compared to the reference set. Furthermore, all data we used provided almost identical channel characteristics, which often is not the case when clean speech and noise samples are mixed. The results of our exemplary evaluations on acoustic modelling and robust automatic speech recognition show, that the simulation of the environmental conditions shifts the acoustic models trained on this data away from clean speech closer towards the realistic noisy data. But even for our very controlled and almost ideal conditions for a simulation, the results point out that simulated noisy utterances are still far away from realistic noisy speech. While the tendencies of improvement for both ETSI robust front end and Spectral Subtraction for simulated and realistic noisy speech were comparable, the magnitudes did not match in all cases.

This leads to the conclusion, that the mathematical model used in this and in many other scientific work is probably not sufficient for describing the complexity of generic noisy speech. Considering the Lombard effect as an addition mathematical function, for example, might be one step to improve the results of a simulation.

7. Acknowledgements

This work was kindly supported by the project MediaGrid funded by the Bundesministerium für Bildung und Forschung (BMBF) in Germany.

8. References


<table>
<thead>
<tr>
<th></th>
<th>clean am</th>
<th>simulated am</th>
<th>noisy am</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfcc</td>
<td>31.91</td>
<td>25.96</td>
<td>30.71</td>
</tr>
<tr>
<td></td>
<td>37.50</td>
<td>38.16</td>
<td>35.22</td>
</tr>
<tr>
<td></td>
<td>32.85</td>
<td>30.60</td>
<td>43.06</td>
</tr>
<tr>
<td>sub</td>
<td>49.79</td>
<td>27.93</td>
<td>32.00</td>
</tr>
<tr>
<td></td>
<td>42.87</td>
<td>34.90</td>
<td>34.09</td>
</tr>
<tr>
<td></td>
<td>39.62</td>
<td>29.13</td>
<td>39.14</td>
</tr>
<tr>
<td>etsi</td>
<td>49.01</td>
<td>34.53</td>
<td>35.61</td>
</tr>
<tr>
<td></td>
<td>41.91</td>
<td>38.67</td>
<td>36.83</td>
</tr>
<tr>
<td></td>
<td>38.97</td>
<td>33.72</td>
<td>42.63</td>
</tr>
</tbody>
</table>

Table 2: Phoneme accuracy for different acoustic models (am) and test sets. Plain acoustic models (mfcc), noise reduction by Spectral Subtraction (ssub) and the ETSI robust front end (etsi) are compared.