REVERBERANT SPEECH RECOGNITION: A PHONEME ANALYSIS

Pablo Peso Parada[1], Dushyant Sharma[1], Patrick A. Naylor[2], Toon van Waterschoot[3]

[1]Nuance Communications Inc. Marlow, UK
[2]Dept. of Electrical and Electronic Engineering, Imperial College London, UK
[3]Dept. of Electrical Engineering (ESAT-STADIUS/ETC), KU Leuven, Belgium
{pablo.peso, dushyant.sharma}@nuance.com,
p.naylor@imperial.ac.uk,toon.vanwaterschoot@esat.kuleuven.be

ABSTRACT

We present a phoneme confusion analysis that models the impact of reverberation on automatic speech recognition performance by formulating the problem in a Bayesian framework. Our analysis under reverberant conditions shows the relative robustness to reverberation of each phoneme and also indicates that substitutions and deletions correspond to the most common errors in a phoneme recognition task. Finally, a model is proposed to estimate the confusability of each phoneme depending on the reverberation level which is evaluated using two independent data sets.

Index Terms: phone recognition, reverberation, confusability factor.

1. INTRODUCTION

Reverberant speech is created in confined spaces by multipath sound propagation from source to receiver which creates multiple delayed and attenuated replicas of the original sound [1]. This acoustic distortion significantly decreases automatic speech recognition (ASR) performance in distant-talking scenarios [2] [3]. Hence it is important to analyse this distortion in more detail.

Phoneme intelligibility degradation for humans due to reverberation was investigated in [4] where the authors showed how reverberation degrades human intelligibility and that errors obtained have the same distribution compared to non-reverberant environments. ASR performance also degrades in the presence of reverberation although the behaviour compared to human intelligibility seems to be different: the indicative error rate is higher in ASR compared to human listeners [5] [6]. In [7] the performance of a digit recognizer is analysed for different reverberation levels obtained by carefully modifying the room impulse response (RIR). The authors demonstrated that the first 50 ms of the RIR barely affect ASR performance whereas the remainder of the RIR has a significant detrimental impact. Tsilfidis et al. [8] investigated the reverberation impact on phoneme recognition showing the performance achieved for the different reverberation levels considered.

We propose to analyse the impact of reverberation on phoneme recognition for numerous reverberant conditions, with a special focus on the confusion found between phonemes. This paper shows the ASR robustness of each phoneme for different reverberation levels. Furthermore, a model to estimate the confusability of each phoneme depending on the reverberation level is derived from an analysis of the confusion matrices.

The paper is organised as follows: Section 2 presents the experimental set up used to obtain different results. An analysis of the impact of reverberation on phoneme recognition is described in Section 3. In Section 4 a method is proposed to compute the confusability of a phoneme depending on the reverberation level. The results obtained are detailed in Section 5 and finally, in Section 6 the conclusions are drawn.

2. EXPERIMENTAL SET UP

In all the experiments performed in this paper we use TIMIT database [9]. This database is phonetically tagged and it contains a good phonetic coverage of American English [10] providing a rich contextual phoneme diversity [11]. These characteristics provide an ideal framework to analyse the reverberation impact per phoneme since each of these phonemes appears in many different contexts.

Two different speech recognizer are implemented to analyse the effect of reverberation in phoneme recognition. First, a HTK [12] context-independent GMM-HMM phoneme recognizer (CI-HTK) is trained following the recipe suggested in [8]. Second, we build an alternative context-independent (CI-KALDI) and context-dependent (CD-KALDI) GMM-HMM phone recognizer using Kaldi toolkit and its recipe s5 [13]. In all cases, a single-pass decoding without lattice re-scoring or feature transformation is performed in order

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° ITN-GA-2012-316969 and from the Research Foundation Flanders.

to reduce the computational cost. The motivation for using a phoneme recognition in this analysis is to avoid potential impact of language model or dictionary rules in the recognition performance and therefore analyse the impact of acoustic distortions more specifically.

The whole TIMIT test set, excluding the 2 dialect sentences (SA), is divided into two independent sets: development (ClnDev) and evaluation set (ClnEval). The latter comprises the TIMIT core test (192 utterances) and the former includes the remaining test recordings (1152 utterances). The initial 61 phonemes in both test sets are collapsed into a set of 39 phonemes [14]. ClnDev is convolved with 140 simulated RIRs, which are flat distributed with C_{50} values from -3 dB to 40 dB as shown in [15], to create the reverberant development set (RevDev). The reverberant evaluation set (RevEval) is generated by convolving ClnEval with 28 simulated RIRs spanning the C_{50} interval [-3dB, 40 dB] and with all real impulse responses (72 RIRs) from MARDY database [16]. The resulting reverberant sets, RevDev and RevEval, are approximately 138 hours and 16 hours long respectively, which cover a wide range of reverberant scenarios.

The parameter used to measure the reverberation level is C_{50} [17] as it has been shown to be highly correlated with ASR performance [15] [8]. C_{50} is computed from the RIR as the ratio of energy in the first 50 ms to the energy after 50 ms using as time reference the arrival time of the direct path [18].

3. IMPACT OF REVERBERATION ON ASR PERFORMANCE

In this section we show the performance of phoneme recognition for a broad range of reverberation levels as well as the phoneme misclassification for clean and reverberant environments. The ASR performance is computed in this paper as follows,

$$\text{PER} = \frac{D + I + S}{N} \tag{1}$$

where N is the total number of phones recognized, D is the number of deletions, S is the number of substitutions and I the number of insertions.

The PER achieved with ClnDev and RevDev for different ASR configurations is displayed in table 1, which shows a clear ASR performance reduction due to the presence of reverberation. Figure 1 describes in more detail the relative phoneme error degradation $r\Delta$PER obtained for different reverberation levels following

$$r\Delta\text{PER}(\%) = \frac{\text{PER}_{\text{RevDev}} - \text{PER}_{\text{ClnDev}}}{100 - \text{PER}_{\text{ClnDev}}} \cdot 100. \tag{2}$$

In the case of low reverberation levels (i.e. $C_{50}\approx$40 dB) the performance of the different phoneme recognizer is scarcely affected. However, an increment of the reverberation level clearly leads to a significant degradation which shows the importance of understanding the reverberation impact on ASR.

	CI-HTK	*CI-KALDI*	*CD-KALDI*
ClnDev	40.2%	35.52%	33.59%
RevDev	66.8%	62.28%	59.45%

Table 1. Phoneme error rate achieved with ClnDev and RevDev.

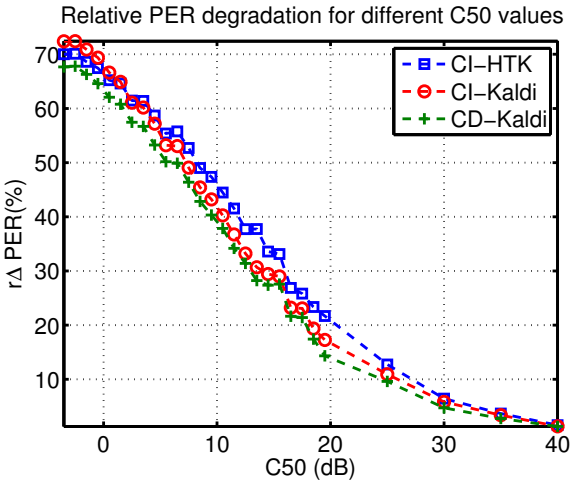

Fig. 1. Relative phoneme error rate degradation $r\Delta PER$ vs. reverberation level C_{50}.

The performance of each phoneme in reverberant environments is prensented in Fig. 2 which plots the confusion matrix obtained with ClnDev and RevDev. These matrices are obtained with the ASR system that provides the best performance in these experiments: CD-KALDI. The matrices are normalized horizontally consequently each cell, for instance row k starting from top and column l starting from left, represents the likelihood of recognizing the phoneme R_l given true phoneme T_k. As a result, the main diagonals in the matrices represent the likelihood of correctly recognizing the given phoneme that is $P(R_k|T_k)$ where k is the phoneme index. The /sil/ label represents a pause. In addition to the 39 phonemes a new label /blk/ representing a blank is included in the matrices to take into account the deletions and insertions. Therefore the last row represents the insertions and the last column the deletions.

Figure 2 provides some insights into the ASR performance under reverberation. Firstly, the correct classification rate per phoneme (main diagonal of the confusion matrices) clearly shows that the correct recognition rate significantly drops when reverberation is present, especially with pauses (/sil/) due to the time smearing of previous phonemes into these low energy gaps. Secondly, the distribution of the insertions (i.e. last row in Fig. 2) is almost equally distributed for all phonemes and is similar under non-reverberant and reverberant conditions. Thirdly, a considerable increase of deletions appear in RevDev as compared to ClnDev owing to time smearing which makes some phonemes to be recognized as the previous one. Finally, some phonemes are confused many

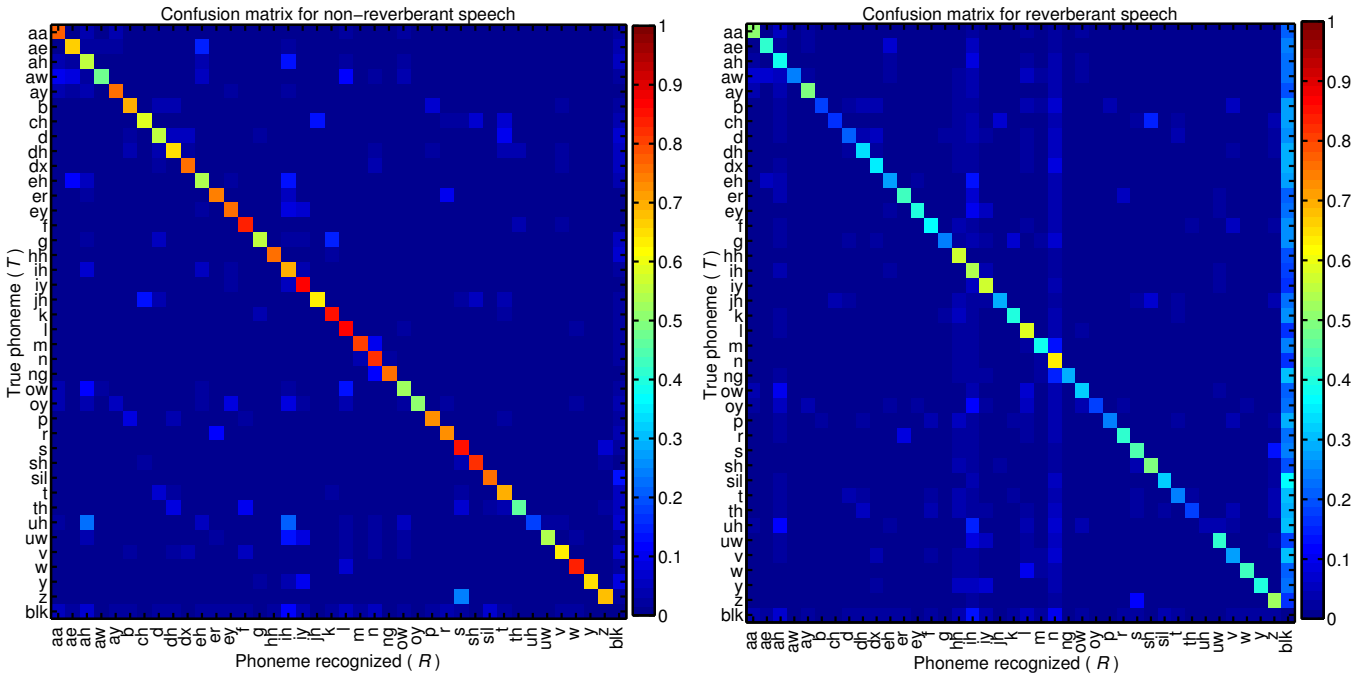

Fig. 2. Phoneme confusion matrix obtained with ClnDev (left plot) and RevDev (right plot).

times, as for example phoneme /hh/, which is shown in the confusion matrix with vertical patterns of high values. This observation is in accordance with the conclusion presented in [19].

Table 2 displays the relative difference (rΔ) of correctly recognized (*H/N*), inserted (*I/N*), deleted (*D/N*) and substituted (*S/N*) phoneme rate between ClnDev and RevDev computed as,

$$r\Delta X = \frac{X_{\text{ClnDev}} - X_{\text{RevDev}}}{X_{\text{ClnDev}}}, \tag{3}$$

where X can be *H/N*, *I/N*, *D/N* or *S/N*.

As expected, the rate of correctly recognized phonemes decreases whereas deletions and substitutions are considerably increased under reverberation. However the insertion rate is slightly reduced. Table 2 indicates that ASR performance degradation is mainly caused by deletions and substitutions.

	rΔ *H/N*	rΔ *I/N*	rΔ *D/N*	rΔ *S/N*
CI HTK	0.41	0.24	-2.37	-0.57
CI Kaldi	0.47	0.66	-2.88	-0.26
CD Kaldi	0.44	0.56	-4.38	-0.54

Table 2. Relative difference of phonemes recognition rates between ClnDev and RevDev.

It is clear that reverberation affects phoneme recognition differently depending on the reverberation level and the phoneme. We aim to model the phoneme errors at the output of the ASR using the confusion matrix which depends on the reverberation level. Such a model would be useful for predicting possible errors or for assigning confidence values to the phonemes derived from the confusability factor. In practice, C_{50} can be blindly estimated applying different methods [20].

4. CONFUSABILITY FACTOR IN A BAYESIAN FRAMEWORK

In this section we present a parameter to measure the confusion between phonemes using a Bayesian framework. Let T_k denote the true phoneme and R_k the recognized phoneme where k represents the phoneme label index. In this paper we consider a set of $P = 39$ phonemes. We propose to compute the confusability factor $\mathcal{CF}(T_k, R_k, \text{C}_{50})$ based on the probability of correctly recognized phoneme index k for a given reverberation level (C_{50}), as follows,

$$\begin{aligned}\mathcal{CF}(T_k, R_k, \text{C}_{50}) &= 1 - p(T_k|R_k, \text{C}_{50}) = \\ &= 1 - \frac{p(R_k|T_k, \text{C}_{50}) \cdot p(T_k)}{\sum_{i=1}^{P+1} p(R_k|T_i, \text{C}_{50}) \cdot p(T_i)},\end{aligned} \tag{4}$$

where the prior probability of the phoneme label T_k is $p(T_k) = \frac{\sum_{i=1}^{P+1} N_{T_k R_i}}{\sum_{i=1}^{P+1} \sum_{j=1}^{P+1} N_{T_i R_j}}$, the likelihood of classifying the phoneme R_k given the phoneme label T_k and the reverberation level C_{50} is $p(R_k|T_k, \text{C}_{50}) = \frac{N_{T_k R_k}}{\sum_{i=1}^{P+1} N_{T_k R_i}}$, and $N_{T_k R_i}$ represents the number of times the phoneme label T_k is classified as R_i for a given C_{50}. It can be shown that the confusability factor presented in (4) can be computed directly from the confusion matrix as follows,

$$\mathcal{CF}(T_k, R_k, \text{C}_{50}) = 1 - \frac{N_{T_k R_k}}{\sum_{l=1}^{P+1} N_{T_l R_k}}. \tag{5}$$

It is worth noting that the phoneme indexes cover the range from 1 to $P+1$ for the purpose of including, in addition to the substitution errors, the insertions and deletions in the computation of the confusability factor.

5. RESULTS

In this section we present the results of the confusability factor for RevDev and we assess repeatability of the results utilizing RevEval.

Figure 3 illustrates the confusability factor presented in (5) with CD-KALDI for each recognized phoneme R_k (rows) at different levels of reverberation as measured using C_{50} (columns). It shows that the phoneme confusion is different for each phoneme and strongly depends on the reverberation level. In all cases, the confusability factor tends to increase when reverberation level increases however the rate of change varies significantly between phonemes. Similar behaviour of the confusability factor was observed for CI-HTK and CI-KALDI.

Combining the confusability factors achieved for each of the 39 phonemes into broad classes of phonemes based on production manner [21] shows that weak fricative phonemes (/th,v,hh,f,dh/) are the most confused class. On the contrary, silence broad phone class, which includes only /sil/ (pause) preserves a low confusability value amongst different reverberation levels. This lower confusion is due to the lack of energy of this phoneme.

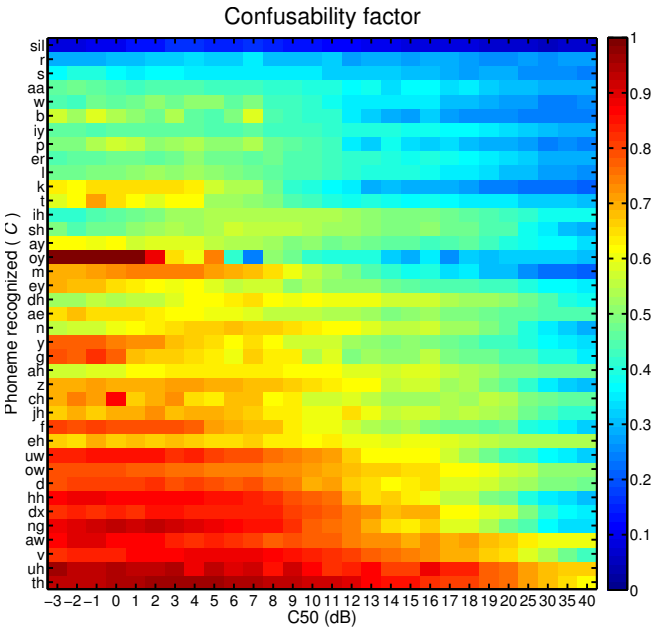

Fig. 3. Confusability factor $\mathcal{CF}(T_k, R_k, \mathrm{C}_{50})$ of the 39 phonemes for CD-KALDI with RevDev.

In order to assess the repeatability of these results we compare $\mathcal{CF}(T_k, R_k, \mathrm{C}_{50})$ computed from unseen RevEval data to a polynomial function fitted to RevDev. Therefore, we fitted a third order polynomial function to $\mathcal{CF}(T_k, R_k, \mathrm{C}_{50})$ for each phoneme k. Hence, one polynomial function is computed for each phoneme or broad phone class and this function depends only on the C_{50} value. The degree of polynomial was chosen such that the function minimizes the root mean square deviation (RMSD) in RevDev. RMSD is computed as follows,

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{n=1}^{N_C}(y_n - x_n)^2}{N_C}}\ \mathrm{dB}, \tag{6}$$

where N_C is the number of different reverberant conditions (i.e. different C_{50} values considered in the reverberant sets), and y_n and x_n are the fitted function output and the $\mathcal{CF}(T_k, R_k, \mathrm{C}_{50})$ respectively for a given phoneme index k and reverberant condition C_{50}.

Table 3 presents the RMSD for RevDev and RevEval of a third order polynomial fitted to RevDev. It shows considerably low deviations for the three ASR configurations. As expected, the error in RevDev is lower because the polynomial function is fitted to this data but the error in RevEval still remains significantly low.

	CI HTK	CI Kaldi	CD Kaldi
RevDev	0.030	0.037	0.035
RevEval	0.060	0.075	0.079

Table 3. Average of RMSD achieved with a third order polynomial.

Since RevEval comprises a completely independent set of RIRs (including mostly real impulses responses) and recordings from RevDev, it is possible to conclude that a set of functions can be used to estimate a confusability factor of the recognized class under completely new reverberant environments. This model depends on C_{50}, apart from the ASR output R_k, which can be estimated employing external methods [20].

6. CONCLUSIONS

In this paper we have analyzed the degradation in phoneme recognition under reverberation with different speech recognition toolkits (HTK and Kaldi). We have demonstrated that, for ASR, phonemes vary in their robustness to reverberation. The confusion matrix presented indicates the ASR robustness of each phoneme to different levels of reverberation. We have also shown that the main errors in our tests are deletions and substitutions. Motivated by these observations, we have designed a metric that characterizes the confusion of recognizing the phoneme in a Bayesian framework. Finally, the results of the experiments have demonstrated that for a strongly reverberant scenario with C50=6 dB, the most robust phoneme is /r/ whereas the most fragile phonemes are the class of weak fricatives (e.g. TIMIT phonetic label /hh,th,v/).

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2005.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] R. Haeb-Umbach and A. Krueger, *Reverberant Speech Recognition*, pp. 251–281, John Wiley & Sons, 2012.

[4] Stanley A. Gelfand and Shlomo Silman, "Effects of small room reverberation upon the recognition of some consonant features," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 22–29, 1979.

[5] B.E.D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, vol. 2, pp. 1259–1262.

[6] Richard P Lippmann, "Speech perception by humans and machines," in *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, 1996, pp. 309–316.

[7] A. Sehr, E. A. P. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[8] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech & Language*, vol. 27, no. 1, pp. 380–395, 2013.

[9] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[10] Lori F Lamel, Robert H Kassel, and Stephanie Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989, vol. 2, pp. 161–170.

[11] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[12] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 1–4.

[14] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[15] P. Peso Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[16] J. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[17] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[18] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fifth edition, 2009.

[19] S. Sandhu and O. Ghitza, "A comparative study of mel cepstra and EIH for phone classification under adverse conditions," in *Proc. IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, 1995, vol. 1, pp. 409–412.

[20] Pablo Peso Parada, Dushyant Sharma, Jose Lainez, Daniel Barreda, Patrick A Naylor, and Toon van Waterschoot, "A quantitative comparison of blind C50 estimators," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2014.

[21] Andrew K. Halberstadt, *Heterogeneous acoustic measurements and multiple classifiers for speech recognition*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, November 1998.