ANALYSIS-BY-SYNTHESIS FEATURE ESTIMATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION USING SPECTRAL MASKS

Michael I Mandel and Arun Narayanan

The Ohio State University, Computer Science and Engineering

{mandelm,narayaar}@cse.osu.edu

ABSTRACT

Spectral masking is a promising method for noise suppression in which regions of the spectrogram that are dominated by noise are attenuated while regions dominated by speech are preserved. It is not clear, however, how best to combine spectral masking with the non-linear processing necessary to compute automatic speech recognition features. We propose an analysis-by-synthesis approach to automatic speech recognition, which, given a spectral mask, poses the estimation of mel frequency cepstral coefficients (MFCCs) of the clean speech as an optimization problem. MFCCs are found that minimize a combination of the distance from the resynthesized clean power spectrum to the regions of the noisy spectrum selected by the mask and the negative log likelihood under an unmodified large vocabulary continuous speech recognizer. In evaluations on the Aurora4 noisy speech recognition task with both ideal and estimated masks, analysis-by-synthesis decreases both word error rates and distances to clean speech as compared to traditional approaches.

Index Terms: analysis-by-synthesis, time-frequency masking, large vocabulary automatic speech recognition, missing data

1. INTRODUCTION

Spectral masking is a technique for suppressing unwanted sound sources in a mixture by applying different attenuations to different time-frequency points in a spectrogram. The ideal binary mask is computed from the original signals before they are mixed and is defined as 1 for all time frequency points where the signal to noise ratio is greater than some threshold and 0 otherwise [1]. Masks can be estimated from observed noisy speech by modeling the speech and/or noise, e.g., [2]. Recent work has shown that both ideal and estimated binary masks can increase intelligibility [1,3–5] and automatic speech recognition accuracy [4, 6, 7].

While spectral masking is performed on short-time Fourier transforms of observations or filterbank outputs, speech recognition typically employs non-linear mappings of these representations such as mel frequency cepstral coefficients (MFCCs) [8]. Because of the non-linearity of this processing, it is difficult to compute features from masked noisy speech that match the features of the original clean speech.

The best performing approaches for recognizing speech from masked representations are spectral imputation [9] and direct masking [10]. Spectral imputation reconstructs missing spectral regions and then extracts standard ASR features using the reconstructed spectrum [9]. While this can work well in some situations, the reconstruction is performed in the spectral domain, so distances between spectra and models are not computed in the domain that is useful to ASR. Direct masking is the direct point-wise multiplication of a spectral mask with the observed spectrum, followed by cepstral feature computation and, most importantly, mean- and variance-normalization of each feature dimension across each utterance. It underestimates the energy of the clean speech in regions of high noise energy, a problem our approach is able to overcome. It was shown in [10] to perform comparably or better than other missing data speech recognition techniques [9, 11].

We pose automatic speech recognition using a time-frequency (TF) mask as an optimization problem over the set of MFCCs that represent an utterance. These MFCCs are optimized both to fit the noisy observation where the mask's spectral gains are high, and to have a high likelihood under a large vocabulary continuous speech recognizer (LVCSR). The weighted Itakura-Saito divergence [12] defines the quality of the fit between a spectrum resynthesized from the MFCCs and the observed spectrum. This divergence has been shown to be a good distortion measure for speech [13, 14] and is related to a number of approaches for estimating cepstra from partial frequency-domain observations [15–17]. Increasing the likelihood under the LVCSR ensures that the resynthesis of hidden spectral regions is speech-like. We show that for both ideal and estimated masks, this approach to estimating MFCCs reduces both the word error rates of transcripts and the distance between the estimate and the original clean speech.

This framework provides a number of benefits. Firstly, it provides a coherent means of combining reliability estimates from spectral masking with the speech knowledge contained in an LVCSR. This LVCSR is in fact the same recognizer used in the final experiments; there is no need to use a different or modified recognizer. As shown in the next section, it is possible to perform the comparison at any stage of analysis/synthesis, so many different mask estimates can be accommodated. And the framework is quite flexible, so additional terms can easily

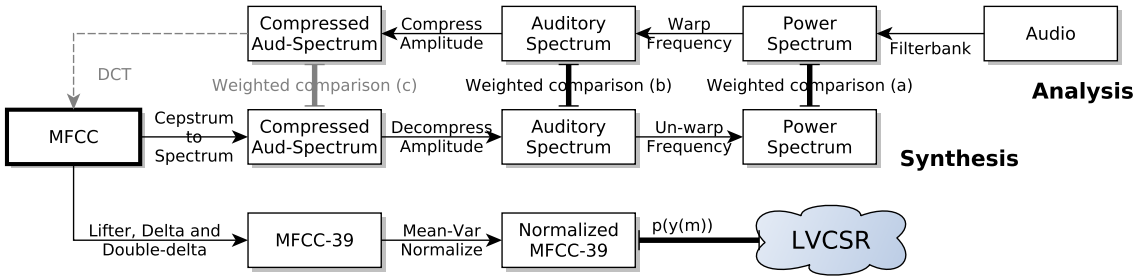

Fig. 1. Flowchart showing the computation of the cost of a particular optimization state (MFCC matrix, shown with a bold outline). The cost combines the distance between synthesis and analysis paths at (a), (b), or (c) using the mask-weighted Itakura-Saito divergence, with the likelihood under a large vocabulary continuous speech recognizer (LVCSR).

be added to the cost function being optimized.

Our approach is similar to spectral imputation [9], except that we reconstruct ASR features directly. It is similar to missing data recognition [18], but allows much more flexibility in the cost function. It is similar to uncertainty propagation [19], but uses an exactly-solved point estimate of the ASR features instead of an approximate distribution over them.

2. ANALYSIS-BY-SYNTHESIS

The analysis-by-synthesis system optimizes the cost function

$$\mathcal{L}(\mathbf{x}; M) = (1-\alpha)\mathcal{L}_I(\mathbf{x}; M) + \alpha\mathcal{L}_H(y(\mathbf{x})) \qquad (1)$$

where $\mathcal{L}_I(\mathbf{x}; M)$ is the Itakura-Saito cost function, $\mathcal{L}_H(y(\mathbf{x}))$ the HMM negative log likelihood, $\mathbf{x}$ is the matrix of MFCCs, $y(\mathbf{x})$ is the matrix of ASR features for the utterance derived from the MFCCs, and M is the time-frequency mask. The parameter α controls the trade-off between matching the observation and the prior model. We found that $\alpha = \frac{1}{3}$ worked well, but that the optimization wasn't particularly sensitive to it. This cost function can be optimized using any unconstrained non-linear method. The two terms in (1) are defined in the following sections along with the closed form gradient calculations used to optimize the HMM log likelihood.

Figure 1 illustrates the computation of the cost function for a given optimization state (MFCC matrix). The top row shows the analysis of audio into various perceptually motivated representations. The middle row shows the synthesis of MFCCs into the same representations, and their comparison to the analysis path. The bottom row shows the conversion of the MFCCs used as the optimization state to the features used in the speech recognizer. Some of these computations are nonlinear and not all are invertible, but all are differentiable.

2.1. Speech recognition features

MFCCs are the standard features for automatic speech recognition. To compute them from audio, the audio is analyzed with a short time Fourier transform, the frequency axis is warped into the more perceptually relevant mel scale, the magnitude is compressed using the log function, and the discrete cosine transform (DCT) is computed across frequency. This process can also be inverted to produce a "filter" corresponding to any MFCC vector. The Itakura-Saito divergence is designed to compare such a smooth synthesis with a less smooth observed spectrum. The analysis path is computed once per utterance. The synthesis path is computed once per optimization iteration using the code from [20]. In order to run the MFCCs through the speech recognizer, they are transformed into full ASR features. After liftering, the delta and double-delta coefficients are computed and each dimension is normalized across each utterance to be zero-mean and unit-variance.

2.2. Masked Itakura-Saito divergence

The analysis path of Figure 1 cannot easily incorporate a spectral mask in such a way that it produces features close to those extracted from the clean speech. We propose instead synthesizing features from MFCCs through the synthesis path so that they match the analysis-path processing up to the comparison point. Depending on the representation for which the mask was computed, synthesized representations can be compared at the points labeled (a), (b), or (c) in the figure, which correspond to measuring the Itakura-Saito divergence between power spectra, between uncompressed auditory spectra, or between compressed auditory spectra (as long as the compressed values are non-negative). We utilize both linear frequency and mel frequency masks in the experiments in Section 3, comparing at points (a) and (b).

Mathematically, for a given MFCC matrix, $\mathbf{x}$, we synthesize a power spectrum matrix, $\tilde{S}_{\omega t}(\mathbf{x})$, and compare it to the noisy observed power spectrum matrix, $S_{\omega t}$, in regions selected by a mask, $M_{\omega t}$, creating the Itakura-Saito cost:

$$\mathcal{L}_I(\mathbf{x}; M_{\omega t}) = D_W\left(S_{\omega t} \| \tilde{S}_{\omega t}(\mathbf{x})\right) \qquad (2)$$

$$= \sum_{\omega,t} W_{\omega t}\left(\frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - \log\frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - 1\right) \qquad (3)$$

The gradient of this quantity with respect to $\mathbf{x}$ is difficult to

derive in closed form, but is relatively inexpensive to estimate numerically because the gradients are independent across time frames. The ability to weight frequencies independently allows us to easily incorporate a spectral mask into this procedure. Specifically, for the experiments in Section 3, we compare ideal binary masks and continuous-valued masks estimated in the mel spectral domain using deep neural networks [2].

When comparing auditory spectra, we use the mask, $M_{\omega t}$, directly as the weights, $W_{\omega t}$ in (2), because the representation is perceptually meaningful. When comparing linear-frequency spectra, however, we multiply $M_{\omega t}$ by a frequency-dependent weighting to better approximate this perceptual importance. To approximate this, we apply a frequency weighting equal to the importance of each linear frequency channel to all warped frequency channels. Mathematically, if a linear-frequency spectrum, S_ω is transformed to a warped frequency spectrum, $S_b = \sum_\omega B_{b\omega} S_\omega$, then the additional weighting that goes into W is (setting DC and Nyquist to 0, which we found to be empirically unreliable)

$$W_{\omega t}(M_{\omega t}) = \begin{cases} M_{\omega t} \sum_b \frac{B_{b\omega}}{\sum_{\omega'} B_{b\omega'}} & 0 < \omega < \frac{f_s}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

2.3. Hidden Markov model likelihood

In order to make the estimated features more speech-like, we add to the cost function the negative log likelihood of the candidate features for the entire utterance under the hidden Markov model (HMM) from an LVCSR trained on clean speech from the Aurora4 corpus [21]. Let $y_{1:T}$ be the matrix of ASR features (mean- and variance-normalized MFCCs, deltas, and double-deltas) derived from $\mathbf{x}_{1:T}$, z_t the hidden states of the HMM, $b_{i,t} \equiv p(y_t \mid z_t = i)$ the probability of the observation at time t under state i, and $a_{ij} \equiv p(z_{t+1} = j \mid z_t = i)$ the transition probability from state i to state j. Then the forward and backward recursions are

$$\alpha_{i,t} \equiv p(y_{1:t}, z_t = i) = b_{i,t} \sum_j a_{ji} \alpha_{j,t-1} \quad (5)$$

$$\beta_{i,t} \equiv p(y_{t+1:T} \mid z_t = i) = \sum_j a_{ij} b_{j,t+1} \beta_{j,t+1} \quad (6)$$

where $\alpha_{i,1} \equiv p(z_1 = i) b_{i,t}$ and $\beta_{i,T} \equiv 1$. The log-likelihood of the data under the HMM is:

$$\mathcal{L}_H(y_{1:T}) \equiv \log p(y_{1:T}) = \log \sum_{\forall z_{1:T}} a_{ij} b_{j,t} \quad (7)$$

$$= \log \sum_i \alpha_{i,t} \beta_{i,t} \quad \forall t \in \{1, \ldots, T\} \quad (8)$$

The gradient of $\mathcal{L}_H$ with respect to a particular y_t is

$$\frac{\partial \mathcal{L}_H(y_{1:T})}{\partial y_t} = \frac{\partial}{\partial y_t} \log \sum_i \alpha_{i,t} \beta_{i,t} \quad (9)$$

$$= p(y_{1:T})^{-1} \sum_i \frac{\alpha_{i,t} \beta_{i,t}}{b_{i,t}} \frac{\partial b_{i,t}}{\partial y_t} \quad (10)$$

Because we are using Gaussian mixture model emissions,

$$b_{i,t} = \sum_k \pi_{ik} \mathcal{N}(y_t \mid \mu_{ik}, \Sigma_{ik}) \quad (11)$$

$$\frac{\partial b_{i,t}}{\partial y_t} = \sum_k \pi_{ik} \mathcal{N}(y_t \mid \mu_{ik}, \Sigma_{ik}) \Sigma_{ik}^{-1} (\mu_{ik} - y_t) \quad (12)$$

Note that we also compute the gradient of $y_{1:T}$ with respect to $\mathbf{x}_{1:T}$. The liftering is preserved in the gradient, the delta and double-delta computations give sums of convolutions across time, the mean normalization goes away, and the variance normalization is preserved. The one approximation we make is to use the variance of the direct-masked features so that the gradient can be computed efficiently.

2.4. Optimization

We optimize the combined cost function (1) using the quasi-Newton BFGS method [22, Section 2-6], where the gradients of $\mathcal{L}_H$ are computed analytically using (10) and (12) and the gradients of $\mathcal{L}_I$ are estimated numerically. The output of this optimization is both the optimal features and the LVCSR recognition result, including the most likely word sequence(s) and state alignments.

Note that additional terms involving any of the intermediate representations of the analysis or synthesis paths can easily be added to the optimization. For example, because the true speech cannot be much louder than the observed mixture, a (soft) hinge loss could be added penalizing a synthesized spectrum that goes too far above the observation.

Computation was performed in a combination of HTK [23] and MATLAB. Instead of computing the HMM gradient over all possible states of the LVCSR model, which would be prohibitive, we approximate this by computing the HMM gradient over the lattice of highest likelihood paths. Specifically, a beam search with a width of 250 nats was used to prune unlikely paths from consideration in constructing the lattice. Additionally, to save computation, this lattice was kept fixed for six iterations of gradient descent and only updated after the sixth. In this way, the recognition and lattice generation could be performed in HTK, with the results loaded into MATLAB for the gradient computation and for reconstructing the ASR features. The alternation between lattice generation and gradient descent was initialized by recognizing the ASR features computed from the direct-masked observation. We found that four iterations of lattice generation (18 iterations of gradient descent) were sufficient to reach the performance ceiling. The full optimization process runs 100 times slower than real time on a single core of an Intel Xeon E5620, 2.40 GHz CPU.

3. ASR EXPERIMENTS

We measure the performance of our feature extraction procedure on speech recognition in the Aurora4, 5000-word closed

Table 1. Word error rates (percentages) on the noisy, matched microphone subset of the Aurora4 test set averaged across all noise types. Direct masking (Direct) vs analysis-by-synthesis (A-by-S). Bold entries are significantly better at a 95% confidence level. A difference of approximately 0.6 is significant.

Mask	Lattice	Direct	A-by-S
Clean		9.54	
Oracle	Clean	14.38	**12.99**
Estimated	Clean	16.18	**13.53**
Oracle	Estimated	14.38	**13.62**
Estimated	Estimated	16.18	**15.31**
Noisy		30.94	

vocabulary task [21]. This dataset consists of speech from the Wall Street Journal (WSJ0) corpus with six different noises added at SNRs randomly selected for each utterance between 5 and 15dB. We used the 7138 clean training utterances and the 996 noisy test utterances (16290 words) without any channel distortion. The recognizer was implemented using HTK [23] with the CMU dictionary for our baseline pronunciations. Tied-state cross-word triphones each modeled as a 3-state HMM with 16 Gaussians per state comprised the acoustic model. A bigram language model is used while decoding.

We compare two masks in the experiments. The first was the ideal binary mask (IBM) defined in the DFT domain using an SNR threshold of 0 dB. The second was an estimated ratio mask (ERM) computed directly in the mel spectral domain using deep neural networks [2]. The recognizer was trained on features extracted from the clean training set using the corresponding feature extraction system. Because of the filtering of the Aurora4 utterances, all versions of the system used masks with frequencies above 7 kHz set to 0.

3.1. Results

The speech recognition results are shown in Figure 2 for each of the six noise conditions and summarized in Table 1. They show that the analysis-by-synthesis approach improves word error rates for both ideal and estimated masks. The table also includes results using lattices from the clean utterances in optimizations for the masked observations, an oracle experiment that places an upper bound on the amount of information that the recognizer could add to the estimation procedure. As can be seen, the clean lattices add a significant amount of additional information, so further gains are possible using this approach if more accurate lattices could be estimated.

A rough measure of reconstructed speech quality is shown in Table 2. Specifically, it shows the Itakura-Saito divergence between the smooth power spectra resynthesized from the MFCC representations and the clean speech power spectrum, averaged across all frames of all mixtures. As shown in [24], the IS divergence is correlated with subjective evaluations of

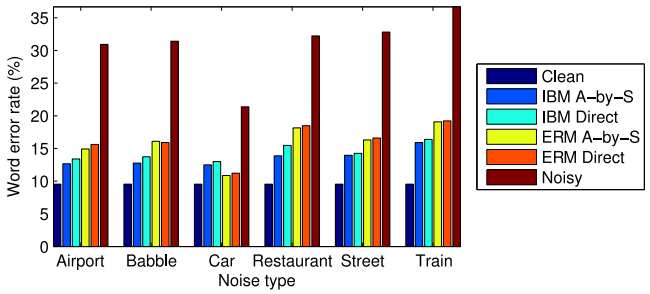

Fig. 2. Word error rate by noise type for direct masking (Direct) and analysis-by-synthesis (A-by-S) with ideal binary masks (IBM) and estimated ratio masks (ERM) using estimated lattices.

Table 2. Itakura-Saito divergence between reconstructed power spectral envelopes and clean power spectra averaged over all frames of all test utterances.

Mask	Lattice	Direct	A-by-S	Δ
Oracle	Clean	273006	272457	-549
Estimated	Clean	276497	275078	-1419
Oracle	Estimated	273006	272506	-500
Estimated	Estimated	276497	275224	-1273
Noisy		272301		

speech quality, but not noise intrusiveness. As can be seen in the Δ column, analysis-by-synthesis reduces the IS divergence, bringing the estimated power spectra closer to the clean signals'. The use of the oracle clean lattice reduces this divergence slightly more than using the estimated lattice, showing that better recognition would improve quality further. Note that the resynthesis from the noisy signal with no masking has the lowest IS divergence with the clean signal, because unprocessed speech has high quality, but high noise intrusiveness.

4. CONCLUSIONS

We have described a new optimization-based analysis-by-synthesis algorithm for extracting automatic speech recognition features from partial spectral observations. The masked Itakura-Saito divergence takes advantage of reliable spectral information while the LVCSR system takes advantage of high-level speech structure. This approach is able to reduce both the word error rates and the distance to the clean speech for both ideal and estimated masks, while providing flexibility to add new information to the optimization in the future.

5. REFERENCES

[1] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., chapter 12, pp. 181–197. Springer US, Boston, 2005.

[2] A. Narayanan and D. L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," Tech. Rep. OSU-CISRC-6/13-TR14, Ohio State University Department of Computer Science and Engineering, 2013.

[3] Douglas S. Brungart, Peter S. Chang, Brian D. Simpson, and Deliang Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[4] Michael I. Mandel, Scott Bressler, Barbara Shinn-Cunningham, and Daniel P. W. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 7, pp. 1872–1883, 2010.

[5] Gibak Kim, Yang Lu, Yi Hu, and Philipos C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, Sept. 2009.

[6] A. Narayanan and D.L. Wang, "On the role of binary mask pattern in automatic speech recognition," in *Proceedings of Interspeech*, 2012.

[7] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. May 2013, pp. 7092–7096, IEEE.

[8] Lawrence Rabiner and Biing H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, second edition, Apr. 1993.

[9] Bhiksha Raj, Michael L. Seltzer, and Richard M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, Sept. 2004.

[10] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and DeLiang Wang, "A direct masking approach to robust ASR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 1993–2005, Oct. 2013.

[11] Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.

[12] Fumitada Itakura and Shuzo Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53, no. 1, pp. 36–43, 1970.

[13] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Sept. 2008.

[14] Robert M. Gray, Andrés Buzo, Augustine H. Gray, and Yasuo Matsuyama, "Distortion measures for speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[15] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 2, pp. 411–423, Feb. 1991.

[16] T. Galas and X. Rodet, "Generalized functional approximation for source-filter system modeling," in *Proc. Eurospeech*, 1991, pp. 1085–1088.

[17] O. Cappe, Jean Laroche, and Éric Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*. Oct. 1995, pp. 213–216, IEEE.

[18] M. Van Segbroeck and H. Van Hamme, "Advances in missing feature techniques for robust Large-Vocabulary continuous speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 123–137, Jan. 2011.

[19] R. F. Astudillo, D. Kolossa, P. Mandelartz, and R. Orglmeister, "An uncertainty propagation approach to robust ASR using the ETSI advanced Front-End," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 824–833, Oct. 2010.

[20] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005.

[21] N Parihar, J Picone, D Pearce, and HG Hirsch, "Performance analysis of the aurora large vocabulary baseline system," in *Proc. Eurospeech*. Citeseer, 2003, pp. 337–340.

[22] R Fletcher, *Practical methods of optimization*, Wiley-Interscience, 1987.

[23] 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.

[24] Yi Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.