Directly comparing the listening strategies of humans and machines

Viet Anh Trinh, Michael Mandel

***Abstract*—Automatic speech recognition (ASR) has reached human performance on many clean speech corpora, but it remains worse than human listeners in noisy environments. This paper investigates whether this difference in performance might be due to a difference in the time-frequency regions that each listener utilizes in making their decisions and how these "important" regions change for ASRs using different acoustic models (AMs) and language models (LMs). We define important regions as time-frequency points in a spectrogram that tend to be audible when the listener correctly recognizes that utterance in noise. The evidence from this study indicates that a neural network AM attends to regions that are more similar to those of humans (capturing certain high-energy regions) than those of a traditional Gaussian mixture model (GMM) AM. Our analysis also shows that the neural network AM has not yet captured all the cues that human listeners utilize, such as certain transitions between silence and high speech energy. We also find that differences in important time-frequency regions tend to track differences in accuracy on specific words in a test sentence, suggesting a connection. Because of this connection, adapting an ASR to attend to the same regions humans use might improve its generalization in noise.**

***Index Terms*—Noise, Speech perception, Sentence recognition, Automatic speech recognition**

I. Introduction

NORMAL-hearing human listeners are remarkably good at understanding speech in noise, much better than ASR systems [1]–[4], even without any grammatical or linguistic information at all [5], [6]. The reasons for these differences, however, are not well understood, and understanding them would very likely directly lead to improvements in ASR noise robustness. Thus, it is reasonable to compare human speech recognition with automatic speech recognition (ASR) to understand the differences between them, why these differences exist, and how the ASR can learn from humans to improve its performance in noise.

We have introduced a method that can reveal the strategy that a human or machine listener uses in recognizing a particular utterance in noise [7]–[9]. By strategy, we mean the combination of time-frequency "regions" that a listener utilizes to recognize a particular utterance when mixed with a particular noise instance in the context of a particular task. In this paper, we compare the importance maps of human listeners to machine listeners. We identify important regions, which are time-frequency points in a spectrogram that tend to be audible when the listener (human or ASR) correctly recognizes an utterance in noise.

Our first experiment (Section V-A) compares human and machine importance maps on the small-vocabulary GRID dataset [10]. Our second experiment (Section V-B) examines the role of the acoustic model on the large-vocabulary AMI dataset [11]. The current paper incorporates and builds upon our previous work [12], which used this technique to analyze human speech perception and a GMM-HMM ASR on the small-vocabulary GRID dataset.

Our work is motivated by approaches from several fields. [3] surveyed human and non-neural-network ASR on several datasets with different vocabulary sizes and concluded that the performance gap between humans and ASR became larger with a harder or noisier test set. [6] focused on analyzing the performance of ASR acoustic models, using a "null grammar" to avoid the influence of a language model. They showed that the WER of the ASR rises much more quickly than that of the humans as the noise level increases.

[5] analyzed phoneme confusions between humans and ASR and showed that a GMM-HMM ASR does not utilize voicing information, which humans do, leading to a high error rate in some cases, for example, recognizing "p" where the actual character is "b". [1] contrasted human and ASR performance in single-channel and multi-channel speech in different noise scenarios. They showed that in a diffuse-noise environment with moving speakers, the ASR requires a 12dB higher signal to noise ratio (SNR) to achieve the same accuracy (50%) as human listeners. [13] evaluated the impact of intrinsic variations in speech on the recognition performance of human and machine listeners. The paper also demonstrated that the SNR needs to be increased by 13 dB for the ASR to achieve human-level performance in a dataset with variation in accent and dialect (a subset of the Oldenburg Logatome dataset [14]).

Additionally, several projects have endeavored to improve ASR noise robustness by building confidence measures of recognition hypotheses based on understanding the errors the recognizer makes and its state when making them [15], [16]. Others have created synthetic data according to various statistical assumptions made in ASR systems [17]–[19], estimating the proportion of errors caused by each assumption.

And others [20] have applied neurophysiological techniques to a deep neural network acoustic model to try to understand its similarities to human speech perception in quiet environments. [20] showed that each neuron in a neural network acoustic model tends to be activated by specific types of phonetic

Viet Anh Trinh is currently at the Department of Computer Science, The Graduate Center, The City University of New York (CUNY),New York, 10016 e-mail:vtrinh@gradcenter.cuny.edu.

Michael Mandel is currently at the Department of Computer and Information Science, Brooklyn College, CUNY, Brooklyn, New York, 11210 and at the Department of Computer Science, The Graduate Center, CUNY e-mail:mim@sci.brooklyn.cuny.edu.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes several example noisy utterances. Contact vtrinh@gradcenter.cuny.edu and mim@sci.brooklyn.cuny.edu for further questions about this work.

features, which is also the case in the human brain, where individual auditory neurons can be selective to different phonetic features [21]. This analysis has focused on clean speech, however, and so does not provide much insight into noise robustness. In addition, responses are averaged across many instances of each phoneme, so cannot provide insight into decisions on individual stimuli or guidance for modifying predictions.

These works mainly focus on comparing the WER performance of humans and ASR or the types of errors that each makes. Our work is different in that we localize the *cause* of these differences in time and frequency by finding where in the spectrogram each listener is paying attention, and how these important regions vary across different listeners, including different acoustic models and language models. Although our analysis is based on a spectrographic representation for visualization purposes, listeners are paying attention to time-frequency portions of the actual speech signal in their auditory representations. By explaining the cause of this disparity, this work could lead to a method to improve ASR performance in noisy conditions. One possible mechanism to achieve this is the Bubble Cooperative Network [22], in which a data augmentation agent is trained to add noise to unimportant regions while simultaneously training the ASR.

There have been several studies on the topic of finding the regions of speech spectrograms that are essential to the task of speech recognition. There is a line of literature focusing on weighting the contribution of different frequency bands to the recognition performance of human listeners [23]–[27].

Another line of research on humans analyzes both frequency and time information, leading to an importance score of every time-frequency point in the spectrogram [28]. They derive an importance map as the weights, W, associated with the noisy spectrogram X in the equation $p(y) = f(X^T W + c)$, where y is a binary label based on the response of a human subject. For example, this target label could have the value 0 or 1 depending on whether the human listener responds that the noisy mixture contains the word "aba" or "ada." The generalized linear model is used to find the weights W that best fit the data. The noisy speech is created by adding Gaussian noise to the clean spectrogram. [28] found that the important region for the task of distinguishing "aba" from "ada" is the second formant transition, which agrees with findings in theoretical phonetics. Our method is different in that the bubble noise technique requires fewer noisy mixtures per utterance than the additive Gaussian noise approach, as their noise has smaller time-frequency modulation. In addition, we compare the importance maps of humans to different ASR systems, including both neural networks and non-neural network systems. [29] applied the bubbles technique to the modulation spectrum domain in audio. They find the value of spectral and temporal modulations, that are vital to general intelligibility in a modulation power spectrum [30] while we focus on time-frequency regions that are important to recognize a particular phoneme in a spectrogram.

Recently, [31] also proposed an approach to analyzing the speech cues using a "bubble" technique. In their approach, the area inside the Gaussian bubble has more noise than the surrounding time-frequency points. In contrast, there is less noise inside the bubble in our method. [32] located the speech cues in time and frequency bands, using truncation in time and low- and high-pass filtering in frequency. As a result of the truncation approach, this technique can only be utilized for the first and last phonemes of an utterance. Our method, however, can find the speech cues of a phoneme at any position, even in a long sentence. [33] showed that human listener recognition error rates increase when the speech cues identified as essential by this method are eliminated while [34] demonstrated that consonant recognition performance is increased when these speech cues are enhanced. Moreover, our task is closely related to the topic of finding a saliency map in computer vision, which aims to find the pixels in an image that are essential for a classifier to make a particular decision [35], [35]–[40]. There are studies that focus on a specific type of classifier, such as convolutional neural networks [35]–[37], which cannot be directly applied to traditional HMM-GMM ASR systems or RNN-based ASR systems. Our current method, however, can apply to any ASR system.

II. Methodology

The core idea of our technique is to measure the intelligibility of a single recording of an utterance mixed with many different instances of noise varying in both time and frequency. Mixtures in which the utterance is intelligible must have revealed a sufficient amount of information from that utterance for the listener to correctly distinguish it from alternatives. Mixtures in which it is not intelligible, must not have revealed sufficient information. Thus time-frequency regions that are frequently audible in intelligible mixtures and inaudible in unintelligible mixtures are likely to represent the location of important cues that the listener is using. By measuring the correlation between audibility of each time-frequency point with the overall intelligibility of the utterance across mixtures, we can compute the importance of each time-frequency point, which we call the time-frequency importance function (TFIF). Details of the method are given in the following sections.

Our method was inspired by the "bubbles" technique in vision [41], which introduced a technique to localize information in pictures of faces that viewers use to classify the gender, identity, and emotions of the face. Our method represented a translation of this approach to the study of auditory perception.

A. Noise process

Each sentence was mixed with many instances of "bubble" noise [7]. This noise was designed to provide glimpses of the speech only in specific time-frequency areas, which we call bubbles. To construct this noise, we began with speech-shaped noise with a signal-to-noise ratio (SNR) of -24 dB, sufficient to make the speech completely unintelligible. The noise was then attenuated in "elliptical" bubbles (more accurately described as jointly parabolic in time and ERB-scale frequency [42]), providing glimpses of the speech in these regions. Within each bubble, the noise was suppressed by up to 80 dB. The bubbles were 350 ms wide at their widest and 7 ERB high at their highest, the smallest values that would avoid introducing audible artifacts. These settings led to a half-width of 90 ms and half-height of 1 ERB. The center points of the bubbles were selected uniformly at random in time and in ERB-scale frequency, except that they were excluded from a 2-ERB buffer at the bottom and top of the frequency scale to avoid edge effects.

B. General task for listeners

In the human listening test, one sentence was selected at random, mixed with bubble noise, and presented to the listener, who then chose a sentence from a list of all of the sentences. Sentence presentation was blocked, so that every block of mixtures used each of the sentences once in a random order. The number of bubbles per sentence controlled the difficulty of the task, and was adapted using the weighted up-down procedure [43] separately for each sentence. When a sentence was correctly identified, the number of bubbles used in its next presentation was reduced, increasing the difficulty of the task, and when it was incorrectly identified, the number of bubbles used in its next presentation was increased, decreasing the difficulty of the task.

In the machine experiment, we trained the ASR according to the standard "recipe" for each dataset. The ASR was then evaluated on test utterances mixed with instances of bubble noise. Instead of using an adaptive scheme as the human listening test, the ASR's noisy test set was created with a fixed number of bubbles per second such that the accuracy of the ASR was approximately 50% on as many words as possible. The ASR task was to output the text given the noisy speech and its accuracy was scored separately for each word in the sentence.

C. Analysis technique

In order to analyze the results, we computed the point-biserial correlation between the dichotomous variable y_{ij}, whether or not the listener correctly identified the jth mixture of the ith utterance, and the continuous variable $A_{ij}(f,t)$, the audibility of time-frequency point (f,t) in the jth mixture of the ith utterance. The intelligibility y_{ij} had value zero if the listener recognized the speech incorrectly and one otherwise. Audibility here is defined as the proportion of attenuation (in dB) applied to the noise at that point, i.e., the depth of the bubble, ranging between 0 for no attenuation (pure noise) and 1 for total attenuation (no noise). This correlation was performed across mixtures, but separately for each time-frequency point for each utterance, leading to a "massively univariate" correlation, denoted $c_i(f,t)$.

$$c_i(f,t) = \frac{m_{i1} - m_{i0}}{s_{in}} \sqrt{\frac{n_{i0} n_{i1}}{(n_{i0} + n_{i1})^2}} \tag{1}$$

where n_{i0} and n_{i1} are the number of incorrectly and correctly identified mixtures ($y_{ij} = 0, 1$) of the ith sentence, respectively, m_{i0} and m_{i1} are the mean audibility in the group of incorrectly and correctly identified noisy mixtures of the ith sentence, respectively, and s_{in} is the standard deviation of all the mixtures ($n_{i0} + n_{i1}$) of the ith utterance. The significance of this correlation was assessed using a two-sided t-test with a test statistic $s_i(f,t)$

$$s_i(f,t) = c_i(f,t) \sqrt{\frac{n_{i0} + n_{i1} - 2}{1 - c_i(f,t)^2}} \tag{2}$$

From the test statistic $s_i(f,t)$, we derived the value of the Student's t cumulative distribution function $F_X(s_i(f,t)) = P(X \leq s_i(f,t))$, where X is a random variable following Student's t-distribution. The two-sided p-value of $p_i(f,t)$ is derived from $F_X(s_i(f,t))$ via

$$p_i(f,t) = \begin{cases} F_X(s_i(f,t)) & \text{if } F_X(s_i(f,t)) \leq 0.5 \\ 1 - F_X(s_i(f,t)) & \text{if } F_X(s_i(f,t)) > 0.5 \end{cases} \tag{3}$$

The resulting p-value for each point and utterance, $p_i(f,t)$, was compared to the significance level of 0.01 to determine if the point-biserial correlation was significantly different from zero.

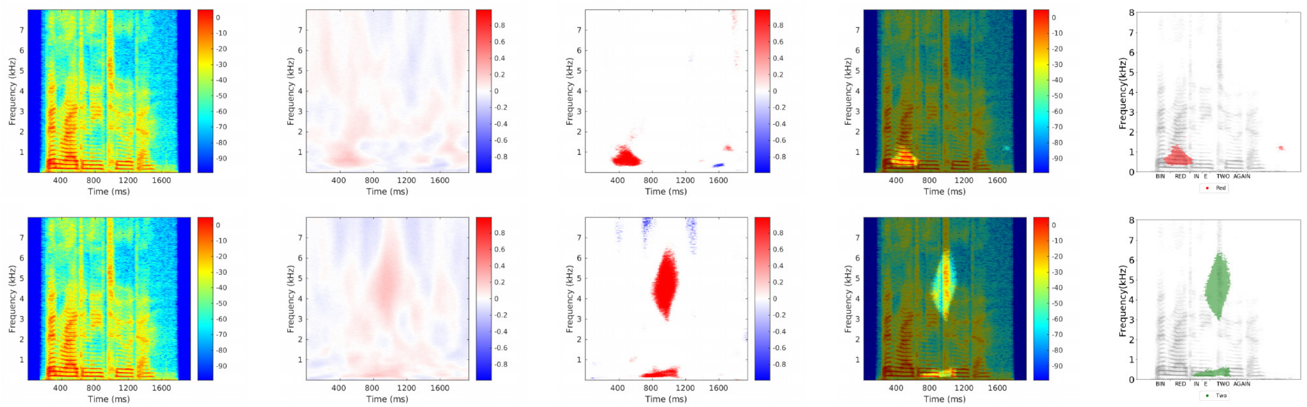
While the task itself was a choice between sentences, these choices could also be analyzed at the word level. Thus the same responses to the same stimuli could be interpreted as having several different meanings for the purposes of our analyses, similarly to the information transmission analysis of [44].

D. Visualization

We also demonstrated various steps leading to our importance map visualization in Figure 1. First, the spectrogram of the clean speech, which is the background of the importance map is shown in Figure 1(a). Note that both rows used the same stimulus, the GRID sentence "Bin red in E two again," (see Section III for details) so show the same spectrogram. Next, in Figure 1(b), we visualize the correlation between the intelligibility of the noisy mixtures and audibility at each time-frequency point with red showing positive correlations and blue showing negative. For this single sentence, we scored the intelligibility of two different words, "red," shown in the top row, and "two," shown in the bottom row. In Figure 1(c), we visualize the value

$$q_i(f,t) = \operatorname{sign}(c_i(f,t)) \exp\left(\frac{-p_i(f,t)}{0.01}\right) \tag{4}$$

which was derived from the significance $p_i(f,t)$ of every time frequency point. In our previous publications, we used the visualization style shown in Figure 1(d), where a false color spectrogram has its lightness (in the HSV color space) set to $0.5 + 0.5 \exp\left(\frac{-p_i(f,t)}{0.01}\right)$ for positive $p_i(f,t)$, so that significant correlations were shown at full lightness and insignificant correlations were shown at half lightness. This only permitted the visualization of a single response per spectrogram. In order to visualize multiple responses per spectrogram, here we introduce the visualization shown in Figure 1(e) of a greyscale spectrogram with a solid color overlaid on it for all points where $q_i(f,t) \geq 0.3679$ which corresponds to a p-value smaller than two-sided significance level 0.01.



(a) Spectrogram (b) Correlation (c) Significance (d) HSV overlay (e) Thresholded signif.

Fig. 1. Various steps leading to the importance maps of two words: “Red” (top row) and “Two” (bottom row) both from the sentence “Bin red in E two again.” (a) Clean speech spectrogram (b) Correlation between the intelligibility of the noisy speech and audibility at each time-frequency point. (c) Signed significance of the correlation, $q(f,t)$. (d) Importance visualization overlaying $q(f,t)$ as brightness in the HSV color space on a pseudocolor spectrogram. (e) Thresholded signed significance overlaid on the clean spectrogram (in grey). Subsequent figures will use visualization (e) which permits the importance of multiple words to be shown on as single spectrogram in different colors.

TABLE I
SENTENCES SELECTED FROM THE GRID CORPUS. ALL SENTENCES WERE SPOKEN BY TALKER 16, A FEMALE.

ID	Verb	Color	Prep	Let	Num	Adv
BBIKZA	“Bin	blue	in	K	zero	again.”
BGIL8A	“Bin	green	in	L	eight	again.”
BRIE2A	“Bin	red	in	E	two	again.”
BRIK6A	“Bin	red	in	K	six	again.”
BRIRZA	“Bin	red	in	R	zero	again.”
BWIE8A	“Bin	white	in	E	eight	again.”
BWIL2A	“Bin	white	in	L	two	again.”
BWIR6A	“Bin	white	in	R	six	again.”

III. DATASETS

A. GRID

The first dataset that we utilized is the GRID corpus [10]. The corpus consists of six-word sentences of the form: $\langle verb \rangle$ $\langle color \rangle$ $\langle preposition \rangle$ $\langle letter \rangle$ $\langle number \rangle$ $\langle adverb \rangle$, such as, “Bin blue in K zero again.” Each position in the sentence has a fixed number of possible entries: 25 letters (excluding W), 10 digits (including “Zero”), and four words in each of the other positions. Each of 34 talkers recorded 1000 sentences, covering all combinations of colors, letters, and digits, and half of the combinations of the other three words. These talkers represent a wide variety of regional British accents.

We selected this corpus because it facilitates both human and ASR experiments. For human experiments, the corpus provides low predictability from one word to another in the sentences. Thus testing the identification of one sentence in noise to a large extent tests the identification of each of the words in it individually in parallel. The words also provide a good balance of phonetic material. One downside of the corpus for our purposes is that the talkers are British and our listeners are American, making the task slightly more difficult and less natural than if it had been recorded by American talkers.

For ASR experiments, GRID provides a large training corpus for building recognizers. This combined with a small vocabulary (50 words total) makes acoustic models easy to train. In addition, there is a baseline recognizer for the challenge distributed with the Kaldi speech recognition toolkit [45], which we utilized. One downside of using the GRID corpus for our ASR experiments is that this baseline system does not include the use of a deep neural network acoustic model.

From these 34,000 utterances, we selected eight to use in the listening tests. Our goal in selecting these sentences was that words in each position be as balanced as possible and as independent as possible from words in the other positions. There was no set of sentences from a single talker in GRID that perfectly satisfied these characteristics, so we selected the set that came as close as possible. The set size of eight sentences was chosen so that an individual subject could perform the entire experiment in a single listening session. The correlational analysis described in Section II-C requires approximately 200 mixtures of each utterance to obtain acceptable levels of significance. These 200 mixtures will take a human 10–15 minutes to listen to and label. Thus an experiment utilizing eight utterances should take 80–120 minutes, a rather long single-session listening test. To minimize the predictability of individual words, we sought a set of sentences where each position in the sentence had a uniform distribution over candidate words. To maximize the independence between adjacent words, we sought a set of sentences with all possible combinations of these words. It turned out not to be possible to find a set with all possible combinations because we found that all sentences differed by at least two words from one another. We thus identified the most confusable set of sentences from a single talker with a distribution over individual words as uniform as possible.

This selection process resulted in the sentences listed in Table I, spoken by talker 16, a female. As can be seen in the table, all of the sentences share the same verb, preposition, and adverb. There are four letters, each appearing in two sentences, and four number, each also appearing in two sentences, both meeting our goals for balance. There are also four colors, but two of them appear in a single sentence each and two appear in three sentences each, not meeting the balance goal, but allowing the words in the other positions to do so.

B. AMI

The AMI dataset is a collection of meeting recordings with 100 hours of speech in English, although the speakers are mainly non-native. The audio is recorded in three different rooms with dissimilar acoustic characteristics. There are several microphone settings such as Individual Headset Microphones (IHM), Single Distant Microphone (SDM), and Multiple Distant Microphones (MDM), and we use the IHM setting for our experiments. The training set is the standard training set of the AMI IHM recipe in Kaldi [45]. There are two AMI test sets: the first one is the standard AMI test set and the second one, the bubble test set, comprises 1000 bubble noise mixtures at -10 dB and 80 bubbles per second created with the single sentence IB4010_H00_FIE038_0022314_0022648, "And you pick up on things that you didn't really notice the first time around.". The SNR and number of bubbles per second are tuned so that the WER is approximately 50% for as many words as possible. The above sentence is chosen because the ASR achieves 100% accuracy on recognizing the words when given the original clean audio.

IV. ASR SYSTEMS

ASR is a very large field and we only review here the very basics necessary to understand our subsequent experiments. The interested reader is directed to various books on the topic [46]–[48]. The ASR task is to convert speech to text. It is made up of several sub-systems that operate in relative isolation from one another. In addition, there is currently an emerging class of end-to-end ASR where the entire system is trained together and these boundaries become less clear. The feature extractor converts the raw audio signal into an intermediate representation of acoustic features. The acoustic model (AM) evaluates the probability of a given phoneme for a given acoustic feature vector. The pronunciation model maps words to sequences of phonemes. Finally, the language model (LM) assigns probabilities to word sequences. By combining these models with an efficient search procedure, the sequence of words with the highest probability given an acoustic utterance can be found. Our experiments evaluate different choices of models for AM and LM, so we discuss them here along with related techniques.

A. GMM acoustic model

For many years, ASR systems used Gaussian mixture model (GMM) acoustic models. The AM can produce probabilities for monophones or triphones, in which case each phoneme is modeled in the context of the phonemes (or phoneme types) before and after it. Expectation maximization is used to find the maximum likelihood parameters for the GMM. Some additional techniques are employed to facilitate modeling of the features with GMMs. Linear Discriminant Analysis (LDA) is used to reduce the dimensionality of the features in accordance with their utility in discriminating phonemes. LDA not only finds the axes that maximize the variance of the data projection on them (similar to PCA), but also maximize the separation among different classes. Maximum Likelihood Linear Transform (MLLT) is used to reduce the variation between the speech of different speakers. It allows the ASR system to better generalize to unseen speakers.

The GMM AM system used on CHiME-2 was based on the Kaldi baseline for the first CHiME challenge [49] which was also the baseline for the second CHiME challenge, track 1 [50]. There are 2 GMM AM systems used on AMI: the first is a GMM-HMM triphone system using MFCCs together with Linear Discriminant Analysis and Maximum Likelihood Linear Transformation (LDA-MLLT), denoted "tri2b." The second adds Speaker Adaptive Training (SAT) [51] and is denoted "tri3.". Both are default systems from Kaldi.

B. Neural network acoustic model

Recent state of the art results have used acoustic models based on time-delay neural networks (TDNN) or a combination of time-delay neural networks and long short-term memory networks (TDNN-LSTM). A TDNN is feed-forward, but utilizes a sort of dilated convolution to provide an increased temporal context to higher layers of the model. TDNNs not only capture long term dependencies like recurrent neural networks, but also have small training times on par with simple feed-forward networks.

The neural network acoustic model used for CHiME-2 is a time delay neural network (TDNN) [52] following the CHiME-5 baseline recipe in Kaldi ported to the CHiME-2 track 1 recipe. The input to the TDNN is a 140-dimensional (140-D) vector, which is the concatenation of a 40-D "high-resolution" MFCC feature (calculated from a 40-D mel filterbank) and a 100-D iVector [52]. The neural network-based acoustic model for AMI is the default hybrid TDNN-HMM chain structure in the 5b Kaldi recipe and uses the same features. In addition, we also analyzed TDNN-LSTM acoustic models with and without dropout [53]. In the default configuration, some layers are projected LSTMs [54], specifically layers 4, 7, and 10, while other layers are TDNNs. Dropout is applied to the i, f, and o gates of the LSTMs. To the best of our knowledge, our Kaldi system for AMI is representative of the state of the art (SOTA). It achieves a WER of 19.3 on the test set. For example, a recent paper [55], achieves a WER of 22.02. The current SOTA on CHIME-2 track 1 is [56], where a speech enhancement system is used on the input to the ASR. We chose not to include speech enhancement at the moment to evaluate the importance maps of the ASR itself. Analyzing a combined speech enhancement ASR system is left for future work. As such, our systems for CHIME-2 track 1 are the best systems provided with Kaldi, which tries to be close to SOTA while striving to maintain generality.

C. n-gram language models

The standard n-gram language model makes a Markov assumption that the probability of a word only depends on the previous $n-1$ words. This probability is derived using maximum likelihood estimation as:

$$p(w_i|w_1,\ldots,w_{i-1}) \approx p(w_i|w_{i-n+1},\ldots,w_{i-1}) \quad (5)$$

$$= \frac{C(w_{i-n+1},\ldots,w_i)}{C(w_{i-n+1},\ldots,w_{i-1})} \quad (6)$$

where $C(\cdot)$ is the count of how many times a word sequence appears in the training corpus. We use the SRI Language Modeling Toolkit (SRILM) [57] as the n-gram language modeling tool with Kaldi. SRILM supports different kinds of smoothing algorithms such as Kneser-Ney [58] and Jelinek-Mercer [59] to deal with the problem of test n-grams unseen in the training corpus. The SRILM perplexity on the AMI test set is 79.7.

D. RNN language models

The n-gram technique is simple, but since the size of the model grows exponentially with n, it can not capture long-term dependencies, and the count tensor is sparse. Another way to build a language model is to use a neural network to estimate the probability of a word given its previous context. A model that is suitable for this task is the recurrent neural network, which has the ability to capture long-range dependencies better than the n-gram technique. In this paper, we use the RNN language model (RNNLM) from Kaldi [60]. The Kaldi RNNLM is optimized by applying importance sampling-based methods and utilizing unnormalized probabilities during training [60]. In addition, it uses subword information to improve the word-embedding representation for out-of-vocabulary and rare words. The RNNLM is trained on the Fisher and AMI corpora following the Kaldi AMI recipe "s5b". The RNNLM perplexity on the AMI test set is 54.2.

V. Experiments

A. Experiment 1: Importance maps on the GRID dataset

Our first experiment compares the importance maps of human listeners with machine listeners on recognizing the eight GRID sentences listed in Table I. The machine listeners include a GMM acoustic model and a TDNN acoustic model using either an eight-sentence language model or the full 64,000-sentence GRID language model.

1) Human importance maps: In the human listening test, one sentence was selected at random, mixed with bubble noise, and presented to the listener, who then selected one of the eight possibilities. As mentioned in Section II-B, the difficulty was adjusted online using the weighted up-down procedure [43]. It was applied separately for each sentence starting at 30 bubbles per sentence. When a sentence was correctly identified, the number of bubbles used in its next presentation was reduced by 2% and when it was incorrectly identified, the number of bubbles used in its next presentation was increased by 2.3%. This asymmetry leads the procedure to converge to the number of bubbles per sentence that allows the listener to correctly identify 56.3% of the mixtures, half way between chance and perfect performance. This procedure resulted in a final bubble rate of 18–24 bubbles per sentence, varying by listener and utterance.

The human listening test was performed over headphones via a MATLAB interface. Subjects consisted of one expert listener, who labeled 1600 mixtures and was familiar with bubble noise, and three naïve listeners, who together labeled another 1600 mixtures (401, 562, and 639 mixtures each) and had never heard bubble noise before. All were native speakers of American English. Subjects were allowed to familiarize themselves with the clean utterances and the task for 5 minutes before noise was added. They were allowed to adjust presentation volume to a comfortable level, listen to each mixture as many times as they wanted, take breaks regularly, and end their participation whenever they wanted. Feedback was provided only at the end of the training period and no feedback was provided during the experiment itself. The listeners each spent approximately the same amount of time on the test, the expert being the fastest.

Because the bubble method can analyze importance at the word-level, in case the ASR does not recognize all the words in a specific sentence, we can still use the correctly identified words for analysis. Thus, we visualize the importance map using both the sentence-level and the word-level "correctness" when we analyze the human listening test.

2) Machine importance maps: When measuring importance maps for ASR systems on the GRID dataset, we examine the role of the acoustic model and language model. First, we compare two different acoustic models on this task, one based on a neural network AM and one on a non-neural network AM. The non-neural network ASR system used in these experiments was based on the Kaldi baseline for the first CHiME challenge [49] which was also the baseline for the second CHiME challenge, track 1 [50]. The training data consisted of speech from the GRID corpus mixed with various noises recorded in a household environment. The recognizer used a Gaussian mixture model (GMM) front end operating on mel-frequency cepstral coefficients (MFCCs) predicting clustered triphone states. The MFCCs were transformed using linear discriminant analysis of nine consecutive frames followed by a global maximum likelihood linear transformation [61].

These GMMs were trained on the training data from the second CHiME challenge, track 1, which consists of 17,000 noisy utterances, 500 from each talker. Our training excluded utterances from talker 16, the one used in the listening test.

Several modifications to the decoding parameters of the GMM AM were necessary to perform the same listening task as the human listeners. First, we modified the grammar to consist of only the eight test sentences as eight parallel paths from the start state to the end state. After doing so, we needed to modify the weights on each of the sentence paths in the grammar to achieve approximate parity in the frequency with which each sentence was selected. This required placing a large penalty (52 nats, where a nat is a unit of information like a bit, but using the natural logarithm) on selecting BWIE8A, moderate penalties (46, 43, 39, and 36 nats) on selecting BGIL8A, BRIE2A, BRIRZA, and BWIL2A, respectively, and low penalties (25, 16 and 0 nats) on selecting BBIKZA, BWIR6A and BRIK6A. Apparently in bubble noise the recognizer was particularly unlikely to select the utterances containing the word "Six." We also found that with the default settings, many sentences' transcripts ended before the final state due to beam search starvation. Because this is a small-vocabulary task, we were able to increase the width of the beam to 200 nats to eliminate this issue, presumably by exhaustively exploring all paths.

As expected, in order to correctly identify 50% of the sentences, the ASR could only tolerate much milder bubble noise than the human listeners. We performed several searches across the number of bubbles per sentence and SNR to identify a good operating point. One advantage of the machine listener over the human listeners is that it can listen faster than real time and has an unlimited attention span. We thus utilized 400 mixtures per utterance with it.

The neural network acoustic model uses a time delay neural network (TDNN) [52] following the CHiME-5 baseline recipe in Kaldi ported to the CHiME-2 track 1 recipe. The input of the TDNN is a 140-dimensional vector, which is the concatenation of a 40-dimensional traditional MFCC-based feature and a 100-demensional iVector [52] We use penalties of 22, 25, 13, 0, 39, 24, 18 and 33 nats for BBIKZA, BGIL8A, BRIE2A, BRIK6A, BRIRZA, BWIE8A, BWIL2A, BWIR6A.

Second, we examine the role of the language model on the importance map. We compare two language models: the first contains only the 8 sentences listed in Table III (TDNN-8LM) and the second is the full GRID language model (TDNN-64kLM) containing 64,000 sentences. The TDNN-8LM perplexity is 2.1 while the TDNN-64kLM perplexity is 6.3. Since the words are not entirely independent, the 8-sentence language model could take advantage of this while the 64k LM could not. The acoustic model is fixed to be the TDNN for both language models. The TDNN-8LM ASR system is analyzed using bubble noise with a baseline SNR of -9 dB while the TDNN-64kLM ASR uses 0 dB so that each achieves an accuracy around 50%.

B. Experiment 2: Importance maps on the AMI dataset

In the second experiment, we examine how different acoustic models and language models change ASR importance maps on the AMI dataset.

First, we compare several acoustic models on the AMI IHM dataset: a conventional Gaussian mixture model (GMM), a time delay neural network (TDNN), a combination of time delay neural network and long short-term memory network (TDNN-LSTM), and a TDNN-LSTM with dropout (LSTMd). These systems are trained on the standard AMI IHM dataset and tested on the two AMI test sets described in Section III.

The non-neural network setting is a GMM-HMM baseline in the Kaldi AMI recipe. We follow the Kaldi script to train a sequence of systems, in which alignments from one system are used to train the next system.

Two non-neural network systems are evaluated: the first is a GMM-HMM triphone system using MFCCs together with Linear Discriminant Analysis and Maximum Likelihood Linear Transformation (LDA-MLLT), denoted "tri2b." The second adds Speaker Adaptive Training (SAT) [51] and is denoted "tri3."

The neural network-based acoustic model is the default hybrid TDNN-HMM chain structure in Kaldi AMI recipe 5b. In addition, we also analyzed TDNN-LSTM acoustic models with and without dropout [53]. In the default configuration, some layers are projected LSTMs [54], specifically layers 4, 7, and 10, while other layers are TDNNs. Dropout is applied to the i, f, and o gates of the LSTMs.

Second, the effect of the language model on the importance map is investigated. We compare the standard n-gram-based SRILM [57] language model with two RNN language models. Both use the same TDNN acoustic model. The RNN language model is better at capturing long term dependencies, but has a longer training time than an n-gram model. We use the Kaldi RNNLM in two distinct ways: for rescoring lattices using a 4-gram approximation to the history [60] and to rescore the top 50 transcripts. Rather than using only the RNN language model, it is combined with SRILM predictions with each given equal weight (0.5). We ported part of this RNNLM from the Switchboard recipe to the AMI recipe.

In general, we observe that GMM acoustic models tend to focus on areas of the spectrogram that contain little speech energy in a wide range of frequency bins while the neural network AMs pay attention to the high-energy, low-frequency regions. We also examine the role of the language model (LM) and show that the importance map of an ASR with a neural LM does not differ much from those of traditional n-gram backoff LMs on the particular sentences studied here.

VI. Results and discussion

A. Experiment 1: Importance map on the GRID dataset

1) Human results and discussion: Time-frequency importance functions derived from the human responses are shown in the top row of Figure 2. It can be seen that the human listeners are attending to time-frequency locations in the spectrogram corresponding to various speech cues. These cues include the initial glides of "white" and "red", the initial stop burst of "two", and the initial sibilance of "six". These results are consistent across different productions of the same word. They also follow the well-established cues of speech production for these words [62] and agree with other analyses of cues for speech perception of individual tokens [32]. The identified regions, however, only include a subset of the distinctive features that might be expected. For example, the final sibilance in "six" does not appear to be utilized consistently, even though it is nearly as energetic as the initial sibilance in BWIR6A. Nor is the initial sibilance in "zero" utilized, although low frequency information in the /z/ does appear to be utilized.

Another interesting feature of these results is that different "correctness" signals (different colors in Figure 2) show correlations with different time-frequency regions of the utterances. This analysis is possible because of the use of sentence stimuli, in contrast to previous auditory bubbles experiments, which employed isolated words [7]–[9]. For word-level correctness, these correlations generally appear in the spectro-temporal regions of the word in question, an effect that is very noticeable in all of the "red" and "white" sentences, but especially for the sentences BRIE2A, BRIRZA, and BWIL2A. For example, in BRIRZA, when the correctness of identifying the color word "red" is considered, the importance is high in the region of the second formant transition of the /r/ in "red". When correctness for the letter word "R" is considered instead, the importance shifts to the second formant transition into and during that word. And when correctness for the number word "Zero" is considered, the importance shifts to the first formant of the initial /z/.

Frequently, however, the importance for one word includes regions of other words. These importance overlaps could be a

TABLE II
CONFUSION MATRICES FOR HUMAN AND ASR WITH TDNN ACOUSTIC MODEL IN EXPERIMENT 1. NOTE THAT THEY ARE RESPONDING TO DIFFERENT TOKENS, WITH SIGNIFICANTLY MORE NOISE IN THE MIXTURES PRESENTED TO THE HUMANS. SENTENCES ARE ABBREVIATED HERE TO THREE CHARACTERS: COLOR, LETTER, NUMBER.

	Human response									ASR response								
True ID	BKZ	GL8	RE2	RK6	RRZ	WE8	WL2	WR6	Sum	BKZ	GL8	RE2	RK6	RRZ	WE8	WL2	WR6	Sum
BKZ	256	30	18	26	43	12	6	9	400	245	42	15	10	28	23	13	24	400
GL8	23	273	16	12	6	39	23	8	400	32	255	7	19	17	23	23	24	400
RE2	4	6	229	41	52	19	42	7	400	12	19	204	67	19	30	26	23	400
RK6	6	6	28	269	21	27	10	33	400	15	11	114	189	13	14	34	10	400
RRZ	10	8	33	20	272	13	22	22	400	26	8	69	62	176	8	18	33	400
WE8	4	15	24	12	15	242	39	50	401	20	35	21	9	12	200	45	58	400
WL2	7	7	38	12	20	27	253	35	399	11	21	26	16	12	63	179	72	400
WR6	5	8	7	54	10	23	42	250	399	17	18	12	64	31	61	94	103	400
Sum	315	353	393	446	439	402	437	414	3199	378	409	468	436	308	422	432	347	3200

result of contextual effects, if pronouncing one word affects the pronunciation of another, or of the task itself, in which recognizing one word can aid in recognizing other words in the sentence. While these two effects are difficult to distinguish, it appears than many of these overlaps could be explained by the construction of the task. For example, in BWIR6A, the importance regions for correctly recognizing the word "R" include both the formant transition of the /w/ in "White" and the sibilance of the /s/ in "Six". This could be explained by the fact that correctly identifying the words "White" and "Six" uniquely identifies this sentence, regardless of whether "R" was audible. Similarly, for the two sentences with unique colors (BBIKZA and BGIL8A), the importance tends to focus on just the color words, even for correct identification of other word positions. This effect can be explained by the fact that the word "Green" only appears in the sentence BGIL8A. Thus, if a human listener can identify the word "Green," then they can predict the other words in the sentence without even hearing them. That is why the importance maps of other words contain the importance map of word "Green.".

2) Machine results and discussion: Figure 2 shows the importance maps for several recognition systems on multiple words in multiple GRID sentences. The systems are the human listeners, the GMM triphone acoustic model, the TDNN acoustic model with the eight-sentence language model, and the TDNN acoustic model with the full GRID language model. The TDNN has importance maps that are much more similar to the human ones than the GMM system's maps, however, it does not capture all the cues that the human listeners utilize.

The GMM generally attends to regions where there is little speech energy. For example, the GMM's importance map for "zero" in the sentence BRIRZA spans the low energy regions from 4 to 8 kHz, however, the human importance for this word is at a high-energy region under 1 kHz. Similarly, there is importance in the low-energy regions at 6-8 kHz for "E" in BWIE8A. Interestingly, there is also importance between the first two formants in "E" in BRIE2A, a low-frequency region of low energy. The GMM TFIFs appear much less sensitive to the type of correctness under consideration, i.e., the importance maps are similar across different word targets. The GMM importance does include some formant transitions, although it seems to be focusing more on the lack of energy adjacent to the formants as opposed to the high energy of the formants themselves. For example, there are large importance regions before the rising second formant of "White" in all of the sentences that include it. Overall these results suggest that the GMM can correctly identify a word or sentence when the noise that is added to it has a similar spectral profile to the speech itself. The GMM thus appears to be using gaps in the noise very differently from the human listeners. While the humans use gaps to identify speech that is revealed, the GMM uses the general spectral shape of the mixture to identify the speech. This is very likely a result of the GMM using MFCC features, which characterize the gross spectral shapes of the entire mixture, and cannot separate the speech from the noise.

The importance maps of the neural network AM with the 8 sentence language model (third row) are more similar to the humans', in particular sharing some cues in high-energy regions. For example, in BRIK6A of Figure 2, the region below 1 kHz of "six" is important for both the human and the neural network AM, but not for the GMM. Moreover, the region around 2 kHz of "R" in BRIRZA is not important for the GMM, but is important for the other listeners. The neural network AM has not, however, captured all the cues that human listeners utilize, such as some specific transitions from silence to high-energy. For example, human listeners are attentive to such a transition above 3 kHz for "two" in BRIE2A and BWIL2A; for "six" in BWIR6A; and to a transition from high energy to low for "six" in BRIK6A. The neural network AM does not pay attention to these specific cues.

The TDNN-64kLM importance maps (the last row) include many low-energy high-frequency regions such as at 4-8 kHz on the right of "again" in BRIE2A. In addition, the importance map of "six" in BRIK6A is located to the left of "bin" at 5-8 kHz, which is unreasonable. Despite these mistakes, the TDNN-64kLM still shares some similar areas with the human maps, such as the region under 1 kHz in "two" in BRIE2A, under 1 kHz for "six" in both BRIK6A and BWIR6A, and the region at 1-2 kHz in "white" in BWIL2A.

Additionally, we quantified the similarity of machine and human listeners importance maps using the Jaccard index, calculated as

$$S = \frac{1}{|N||P|} \sum_{n \in N} \sum_{p \in P} \frac{H_{np} \cap M_{np}}{H_{np} \cup M_{np}} \tag{7}$$

where S denotes the similarity score, N the set of sentences, P the set of positions, $||$ cardinality of a set, and H_{np} (M_{np})

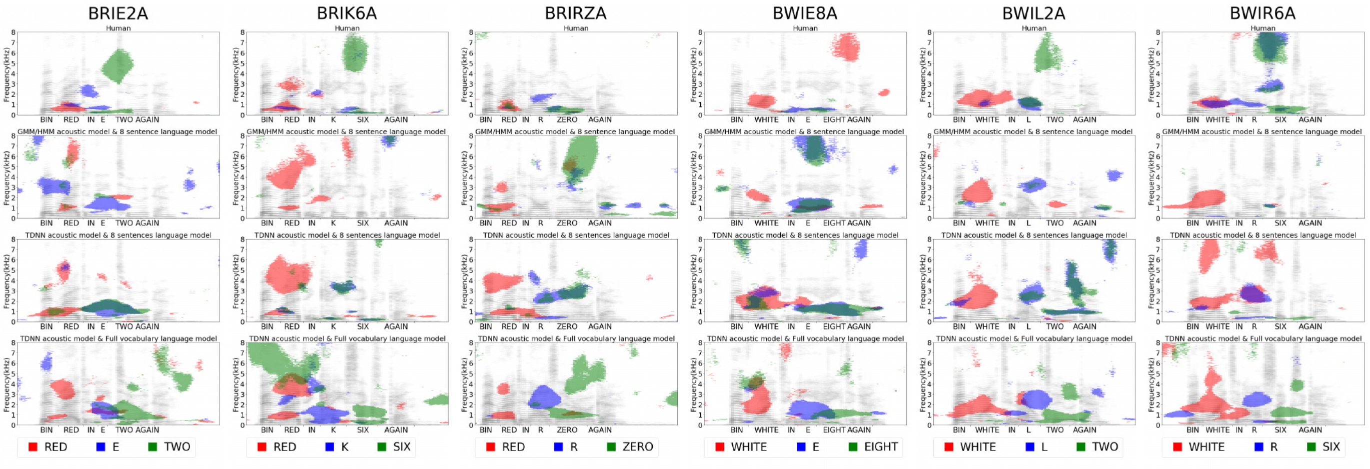


Fig. 2. Importance maps on the GRID dataset for human (top row), GMM acoustic model with 8-sentence language model (second row), TDNN acoustic model with the same LM (third row) and TDNN AM with the 64,000-sentence GRID language model (bottom row).

TABLE III
RECOGNITION ACCURACY (%) OF DIFFERENT COMBINATIONS OF ACOUSTIC MODEL (AM) AND LANGUAGE MODEL (LM) ON THE BUBBLE GRID TEST SET WITH 400 NOISY MIXTURES OF EACH OF THE 8 SENTENCES AT 54 BUBBLES PER SECOND.

AM	LM	SNR	Overall	Letter	Digit
GMM-HMM +LDA+MLLT	8 sents	-9 dB	41.6	44.3	38.8
TDNN	8 sents	-9 dB	55.1	54.2	55.9
GMM-HMM +LDA+MLLT	64K sents	0 dB	23.9	26.8	21.0
TDNN	64K sents	0 dB	43.6	37.2	49.9

the thresholded human (machine) importance maps of word at position p in sentence n.

We choose P as the set of three word positions: second, fourth, and fifth representing the color, letter, and number words. If N is the set of six sentences in Figure 2, then the Jaccard similarity between the human importance map and that of the GMM acoustic model with 8-sentence language model (second row) is 0.045, the TDNN acoustic model with the same LM (third row) is 0.059, and the TDNN AM with the 64,000-sentence GRID language model (bottom row) is 0.089 respectively. When N is the set of all eight sentences in table I, these numbers are quite similar: 0.044, 0.067 and 0.089, respectively. In both cases, we can observe that importance maps for the neural network AM are more similar to humans' than the non-neural network AM's. In addition, somewhat surprisingly, those of the full GRID language model are more similar to the humans' than the task-dependent language model, perhaps because the task dependent language model is too limited to be realistic.

Table III shows the recognition accuracy of these models on the noisy mixtures. The TDNN AM has higher recognition accuracy at 55.1% compared to 41.6% for the GMM. The previous result from [9] showed that the human listeners achieved 63.89% accuracy at -24 dB with 18-24 bubbles per second. Thus the TDNN is both more similar to the human in its importance maps and more noise robust. These results also show that the TDNN-64kLM model (broad search space) has a lower recognition accuracy than the TDNN-8sLM model, at 43.6% compared to 55.1%.

In general, the human experiments highlight the fact that not all speech energy is equally important (e.g., transitions between silence and speech energy are also essential). The non-neural acoustic model (AM), in contrast, did not utilize these cues. Instead, it paid attention to the gross spectral shape of the speech, for instance, a low-energy region around the border of a high-energy area. The neural network acoustic model importance maps contain some high-energy regions similar to human ones but do not include other human cues, such as certain transitions between silence and speech energy.

B. *Experiment 2: Importance maps on the AMI dataset*

The first five rows of Figure 3 show the importance maps of different acoustic models on the AMI test sentence: the GMM-HMM with LDA+MLLT (tri2b), GMM-HMM with LDA+MLLT+SAT (tri3), TDNN-LSTM (without dropout), TDNN-LSTM with dropout, and TDNN. All these acoustic models are combined with the same SRILM language model, except for the sixth row where the TDNN AM is combined with the RNNLM. For visualization purposes, the importance map of each word is shown in a unique color. For instance, the pink salient regions of "really" mainly span from 1kHz to 3kHz in the neural network configurations, while it spans from 1 kHz to 3 kHz and above 3 kHz in the non-neural-network settings. As in the GRID dataset results, correlations in the word-level correctness in the AMI dataset usually emerge in the spectral regions of that word. For example, we see that when the correctness of recognizing "and" is taken into account, the importance is high in the region of the spectrogram that has been force-aligned with "and." When correctness of "really" is considered instead, the importance shifts to the speech energy aligned with "really." Moreover, we can see there some overlap in important regions, thus audibility of one word can be important for recognizing another. This co-occurrence is especially high for "did" and "n't", which are scored separately by the recognizer, but often co-occur.

Generally, the GMM models pay more attention to the low-energy, high-frequency regions than the neural network AMs. For example, for "and," "really," "n't," the tri2b and tri3 models focus on the low-energy high-frequency region at 5 to 8 kHz, while these regions are not important to the TDNN and LSTM.

TABLE IV

Recognition accuracy (%) of individual words in the AMI bubble test sentence. Tri2b: GMM-HMM with LDA+MLLT; Tri3: Tri2+SAT; TDNN-LSTMd is TDNN-LSTM with dropout

AM	LM	"and"	"pick"	"n't"	"really"	"notice"
Tri2b	SRILM	10.6	5.8	30.1	32.3	10.0
Tri3	SRILM	20.3	18.3	56.7	60.5	33.2
TDNN	SRILM	93.6	77.5	96.7	83.8	71.2
TDNN-LSTM	SRILM	92.8	78.7	92.6	86.8	73.9
TDNN-LSTMd	SRILM	89.0	72.2	82.4	72.1	41.3
TDNN	RNNLM	94.3	75.1	96.3	86.3	88.9

Similarly, for "pick," the GMM pays more attention to the low-energy high-frequency region than the neural network AMs.

Table IV shows the accuracy of each model at recognizing several of the words in the noisy test sentence across the 1000 mixtures. The accuracy scores of model tri2b and tri3 on "and," "really," "n't," and "pick" are low compared to the TDNN. Moreover, the TDNN and TDNN-LSTM importance maps look similar to each other as do their scores in Table IV. Thus, we can see that there exists a relationship between the model's importance map and its performance.

Table V shows the recognition accuracy of the different acoustic models on the clean AMI IHM test set and the bubble noise test sentence. It shows that the performance of the neural network acoustic models is more similar on the clean test set and more diverse on the noisy test sentence. For example, the TDNN-LSTM with dropout has a similar WER on the standard AMI test set to that of the TDNN-LSTM without dropout (21.2% vs 21.3%), but has a much higher WER on the noisy test sentence (32.4% vs 22.4%). In Figure 3, the TDNN-LSTMs with and without dropout have similar importance maps for "really," while for "notice" they are different: the importance for "notice" for the model trained with dropout contains an area of low energy at 5.5 to 7 kHz. Interestingly, the two models have a small difference in accuracy on the word with similar importance maps, while the model trained with dropout has a much lower score on "notice" as shown in Table IV. Thus, dropout seems to make the model less robust to noise in our bubble test sentence.

Comparing the results in Table V with the importance maps show in Figure 3 in general, it seems that the recognizers with lower WER utilize more compact regions than those with higher WER. Because the importance maps are identifying points where noise is especially disruptive to recognition, a larger area of such points may indicates a higher sensitivity to noise overall.

Next, we analyze the role of the language model in these importance maps. For the AMI sentence, as can be seen in Figure 3, the importance maps are slightly different between the SRILM (the fifth row) and the RNN language models (the last row). The RNNLM 4-gram lattice rescoring generates similar performance and a similar importance map to RNNLM with n-best rescoring, therefore, we only plot the former. We can see that the importance map of "and" and "really" of the two models shown are quite similar and the recognition accuracy of these words in Table IV is quite similar as well (93.6% compared to 94.3%, 83.8% compared to 86.3%). The importance maps of "notice" are different between the two models, and the recognition accuracies are different as well (71.2% compared to 88.9%).

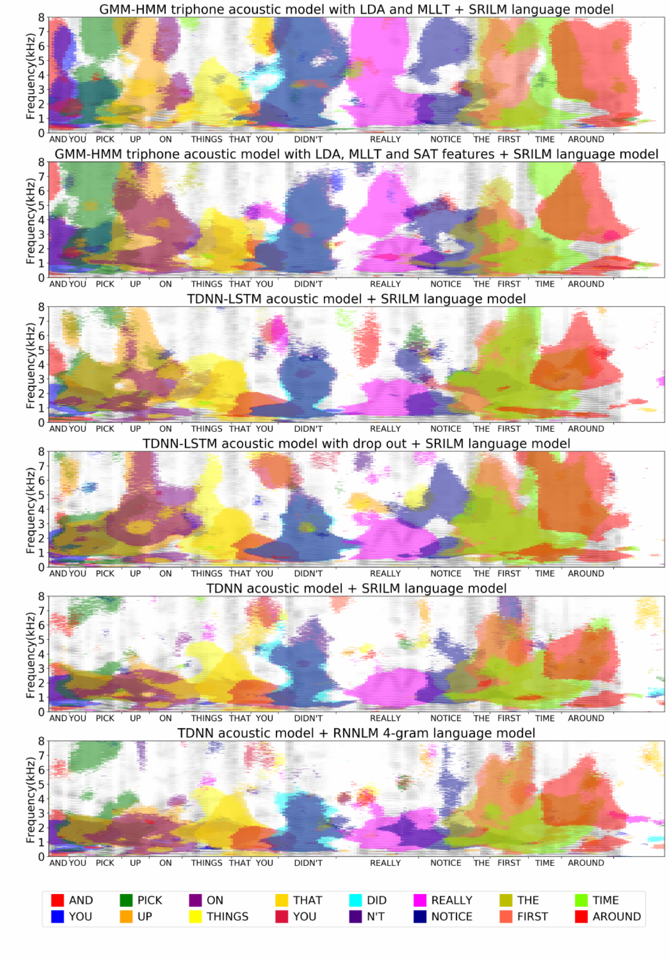


Fig. 3. Importance maps for ASR on AMI test sentence. In order from top: GMM-HMM with LDA+MLLT (tri2b) with SRILM LM, GMM-HMM with LDA+MLLT+SAT (tri3) with SRILM LM, TDNN-LSTM with SRILM LM, TDNN-LSTM with dropout and SRILM LM, TDNN-LSTM with dropout and RNNLN 4-gram language model. The supplementary material contains audio files of example mixtures used to calculate these maps.

TABLE V

Word error rate (%) on the AMI standard test set and the bubble test sentence.

AM	LM	AMI test set	Bubble test sent.
Tri2b	SRILM	39.8	80.1
Tri3	SRILM	35.2	60.1
TDNN	SRILM	21.2	21.2
TDNN-LSTM	SRILM	21.3	22.4
TDNN-LSTMd	SRILM	21.2	32.4
TDNN	RNNLM 4-gram	19.3	20.6
TDNN	RNNLM 50 best	19.6	20.6

VII. Conclusions and Future Work

This paper has described an experiment to directly compare human listeners with automatic speech recognizers in terms of their strategies for recognizing speech in noise. It has shown that in this task, humans focus on time-frequency regions corresponding to formant transitions, stop bursts, and sibilance but a traditional GMM-based acoustic model operating on MFCCs seems to focus on regions of low energy. Additionally, neural network AMs attend to certain high-energy regions similar to those of humans, but do not use all the cues that human listeners make use of, such as certain transitions between silence and high speech energy. While we show large differences in importance and accuracy between different acoustic models on this task, using an RNN language model does not change either one much. These results suggest that the performance of ASR in noisy conditions might be improved by adapting it to pay better attention to the cues used by human listeners, e.g., by using [22]. The paper has also shown that the auditory bubbles technique [7]–[9] can operate just as well on running sentences as isolated words and on machine listeners just as well as human listeners. In the future, it would be interesting to apply the same analysis to end-to-end ASR systems.

Acknowledgments

This material is based upon work supported by the Alfred P Sloan foundation and the National Science Foundation (NSF) under Grants IIS-1618061 and IIS-1750383.

References

[1] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, 2018.

[2] M. P. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, pp. 1562–1573, 2006.

[3] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, Jul. 1997.

[4] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, May 2007.

[5] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *Proc. Workshop on Speech Recognition and Intrinsic Variation*, 2006.

[6] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. EL256–EL261, Mar. 2012.

[7] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *Proc. IEEE WASPAA*, 2013.

[8] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Generalizing time-frequency importance functions across noises, talkers, and phonemes," in *Proc. Interspeech*, 2014.

[9] ——, "Measuring time-frequency importance functions of speech with bubble noise," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2542–2553, 2016. [Online]. Available: http://m.mr-pc.org/work/jasa16.pdf

[10] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[11] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, *et al.*, "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[12] M. I. Mandel, "Directly comparing the listening strategies of humans and machines." in *INTERSPEECH*, 2016, pp. 660–664.

[13] B. T. Meyer, T. Brand, and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 388–403, 2011.

[14] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, "Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[15] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *Proc. IEEE ICASSP*, vol. 1, 1995, pp. 221–224.

[16] L. Chase, "Error-responsive feedback mechanisms for speech recognizers," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1997.

[17] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *Proc. ASRU*. IEEE, Dec. 2011, pp. 71–76.

[18] D. Gillick, S. Wegmann, and L. Gillick, "Discriminative training for speech recognition is compensating for statistical dependence in the HMM framework," in *Proc. IEEE ICASSP*. IEEE, Mar. 2012, pp. 4745–4748.

[19] S. H. K. Parthasarathi, S.-Y. Chang, J. Cohen, N. Morgan, and S. Wegmann, "The blame game in meeting room ASR: An analysis of feature versus model errors in noisy and mismatched conditions," in *Proc. IEEE ICASSP*. IEEE, May 2013, pp. 6758–6762.

[20] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proc. Interspeech*, 2015.

[21] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[22] V. A. Trinh, B. McFee, and M. I. Mandel, "Bubble cooperative networks for identifying important speech cues," *Proc. Interspeech 2018*, pp. 1616–1620, 2018.

[23] K. A. Doherty and C. W. Turner, "Use of the correlational method to estimate a listener's weighting function of speech," *J. Acous. Soc. Am.*, vol. 100, pp. 3768–3773, 1996.

[24] C. W. Turner, B. J. Kwon, C. Tanaka, J. Knapp, J. L. Hubbartt, and K. A. Doherty, "Frequency-weighting functions for broadband speech as estimated by a correlational method," *J. Acous. Soc. Am.*, vol. 104, pp. 1580–1585, 1998.

[25] F. Apoux and S. Bacon, "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," *J. Acous. Soc. Am.*, vol. 116, pp. 1671–1680, 2004.

[26] L. Calandruccio and K. Doherty, "Spectral weighting strategies for sentences measured by a correlational method," *J. Acous. Soc. Am.*, vol. 121, pp. 3827–3836, 2007.

[27] F. Apoux and E. Healy, "Use of a compound approach to derive auditory-filter-wide frequency-importance functions for vowels and consonants," *J. Acous. Soc. Am.*, vol. 132, pp. 1078–1087, 2012.

[28] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception," *Frontiers in Human Neuroscience*, vol. 7, 2013. [Online]. Available: citeulike-article-id:13217478 http://dx.doi.org/10.3389/fnhum.2013.00865

[29] J. H. Venezia, G. Hickok, and V. M. Richards, "Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 1072–1088, 2016.

[30] J. A. Grace, N. Amin, N. C. Singh, and F. E. Theunissen, "Selectivity for conspecific song in the zebra finch auditory forebrain," *Journal of neurophysiology*, vol. 89, no. 1, pp. 472–487, 2003.

[31] L. Varnet, C. Langlet, C. Lorenzi, D. S. Lazard, and C. Micheyl, "High-frequency sensorineural hearing loss alters cue-weighting strategies for discriminating stop consonants in noise," *Trends in hearing*, vol. 23, p. 2331216519886707, 2019.

[32] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, Apr. 2010.

[33] S. Guo, J. Allen, and C. Cole, "Masking effects on perceptual cues in hearing-impaired ears," *The Journal of the Acoustical Society of America*, vol. 145, p. 1797, 2019.

[34] A. Abavisani and M. A. Hasegawa-Johnson, "The role of cue enhancement and frequency fine-tuning in hearing impaired phone recognition," *arXiv preprint arXiv:1908.04751*, 2019.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[37] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[39] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[40] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[41] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261–2271, Aug. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0042-6989(01)00097-9

[42] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[43] C. Kaernbach, "Simple adaptive testing with the weighted up-down method," *Perception & Psychophysics*, vol. 49, no. 3, pp. 227–229, may 1991.

[44] G. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, Mar. 1955.

[45] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[46] F. Jelinek, *Statistical methods for speech recognition.* MIT press, 1997.

[47] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall, 2009.

[48] D. Yu and L. Deng, *AUTOMATIC SPEECH RECOGNITION.* Springer, 2016.

[49] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, may 2013.

[50] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE ICASSP*, May 2013, pp. 126–130.

[51] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

[52] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[53] G. Cheng, V. Peddinti, D. Povey, V. Manohar, S. Khudanpur, and Y. Yan, "An exploration of dropout with lstms," in *Proc. Interspeech*, 2017.

[54] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[55] M. Song, Y. Zhao, S. Wang, and M. Han, "Learning recurrent neural network language models with context-sensitive label smoothing for automatic speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6159–6163.

[56] J. Geiger, F. Weninger, A. Hurmalainen, J. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The tum+ tut+ kul approach to the 2nd chime challenge: Multi-stream asr exploiting blstm networks and sparse nmf," in *Proc. 2nd CHiME Speech Separation and Recognition Challenge held in conjunction with Proc. Int. Congress on Acoustics Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2013, Vancouver, Canada*, 2013.

[57] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[58] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 181–184.

[59] F. Jelinek, "Interpolated estimation of markov source parameters from sparse data," in *Proc. Workshop on Pattern Recognition in Practice, 1980*, 1980.

[60] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on. IEEE*, 2018.

[61] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Tr. SAP*, vol. 7, no. 3, pp. 272–281, may 1999.

[62] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics. MIT Press, 2000.