Utterance Verification for automating the Hearing In Noise Test (HINT)

H. Timothy Bunnell 1, Jason Lilley1, Sigfrid D. Soli2, Ivan Pal3

1 Nemours Biomedical Research, Wilmington DE, USA
2 House Ear Institute, Los Angeles CA, USA
3 Compreval, Inc, Grayslake, IL, USA

bunnell@nemoursresearch.org, lilley@nemoursresearch.org, ssoli@hei.org, ivan.compreval@gmail.com

Abstract

Tests of speech intelligibility play an essential role in many audiological procedures, including diagnostic assessment, verification of hearing aid and cochlear implant fittings, outcome assessment following intervention, and screening of applicants for hearing-critical jobs. The Hearing In Noise Test (HINT) [1] is a speech intelligibility test commonly used for these purposes. A limitation of the HINT, as well as other similar tests, is that they must be administered and scored by a human observer. The present study is an evaluation of a preliminary HMM-based utterance verification system that can be used in place of a human observer to administer HINT.

Index Terms: utterance verification, functional hearing assessment

1. Introduction

The Hearing In Noise Test (HINT) is a means of measuring functional hearing. A listener being tested hears sentences mixed with broadband noise. The listener’s task is to repeat the sentence verbatim to a tester who scores the listeners response as either correct or incorrect. Following each trial in which the listener correctly repeats the sentence that was presented, the noise level is increased so that the next trial will be presented at a less favorable signal to noise ratio (SNR). Following trials on which the listener’s response is not correct, the noise level is reduced and the next trial is presented at a more favorable SNR. Thus, the HINT seeks the SNR at which listeners are able to correctly perceive (and hence repeat) the presented sentence about 50% of the time.

The HINT was developed initially as an American English sentence recognition threshold (SRT) test over 10 years ago[1], based on a Dutch SRT test of similar design [2]. The lists of sentences for the test have been equalized for difficulty to allow their use in adaptive SRT measurements under various test conditions and the reliability and measurement error of the HINT have been established [3]. The HINT technology is available commercially and includes custom software and hardware for automated administration, scoring, interpretation, and report generation. However, the test must be administered by an individual who makes decisions after each response as to the correctness of the response.

Assessment of functional hearing ability is essential to a wide range of non-diagnostic hearing evaluations. For example, functional hearing assessments are used in clinical trials of medical devices, such as cochlear implants, middle ear implants, osseointegrated implants, air conduction hearing aids, as well as various combinations of these devices. The HINT technology has been used, and continues to be used, for baseline and outcome assessments in trials involving all of these devices. Special adaptations of the HINT technology have also been made for cochlear implant assessments[4].

Functional hearing assessment is also an important part of evaluations in occupational health settings. A significant number of jobs, especially those in law enforcement and public safety, include certain tasks for which hearing is essential to the safe and effective performance of the task. The use of the HINT technology for screening in occupational health applications has recently been substantiated in a large laboratory and field study commissioned by the Canadian Department of Fisheries and Oceans (DFO) [5].

While the HINT is widely used in both clinical and occupational settings, one barrier to wider use of the instrument is the need for a trained test administrator to work one-on-one with each individual being tested. The research described here is an initial attempt to remove this barrier by replacing the trained human HINT administrator with an HMM-based utterance verification (UV) system. In this automated version of HINT, each listener response is scored by the UV system as either a correct production of the stimulus sentence, or not, and the SNR of the next trial is adjusted per the output of the UV engine.

Unlike standard ASR applications where input utterances are presumed to be unknown, the task for a UV system is to determine if an input utterance is an acceptable rendition of one specific utterance. Forms of UV systems are implemented in a variety of applications including articulation training software[6, 7], computer-aided language learning [8, 9], and alignment of transcriptions for data-based synthesis[10] to name just a few. The major hurdle in UV is to respond in a way that accurately mirrors the judgments that a well-qualified human listener would make if presented with the same input [11-13]. In the following, we describe the approach taken for UV within the context of HINT administration and present results from an initial live evaluation study.

2. UV Engine Construction

For the present work, a multi-stage approach to utterance verification was used as illustrated in Figure 1. When the UV engine receives a spoken response to a stimulus item, it first applies a restricted automatic speech recognition (rASR) stage for which the only possible utterances are (a) a set of likely non-response responses (e.g., “Don’t know,” “Can’t tell,” “Didn’t get it,”) indicating that the listener is not attempting to reproduce the stimulus item (hereafter, NORESPONSE), (b) a set of known error responses; that is, sentences likely to be uttered in response to misperception of the stimulus sentence (hereafter, ERRESPONSE), and (c) the set of acceptable variants of the stimulus item (hereafter, CRRESPONSE). These three response classes define the Language Model (LM) because they represent the very limited domain of utterances the recognizer expects to receive in response to a specific stimulus item. Separate LMs were constructed for each stimulus sentence using the procedures described below.
The rASR stage of the UV engine returns a single recognized utterance given the LM and the listener’s spoken response. If the utterance class is NORESPONSE or ERRESPONSE, the UV engine returns a final decision of WRONG. However, if the rASR stage returns an utterance from the CRRESPONSE set, a Confidence Measure (CM) is applied to guard against the possibility that an out-of-domain response was a closer match to one of the CRRESPONSE items than to either a NORESPONSE or ERRESPONSE item.

In the following, the methods used to design and train the Phase I UV system are described.

2.1. METHODS

2.1.1. Stimuli

Two sets of speech stimuli were used in training the UV system. The first was the TIMIT speech database (REF), a phonetically diverse set of recordings obtained for 640 speakers recruited from regions throughout the country. Second, four sets of sentences for the HINT test (80 sentences in all) were selected to be recorded under simulated testing conditions designed to maximize the number of ERRESPONSE type utterances that were recorded in response to each HINT stimulus sentence. Thus, these utterances were a mixed set of correct and incorrect responses to HINT sentences when mixed with moderate levels of masking noise.

2.1.2. Subjects

Twenty-five normal-hearing (NH) adults were recruited from the area around the House Ear Institute. All participants were unfamiliar with the HINT test. American English (AE) was the first and primary language for these subjects. No attempt was made to restrict or selectively sample AE dialects.

2.1.3. Data Collection Procedures

Listeners, seated in a sound-dampened chamber, heard each of the 80 HINT sentences presented over headphones. Each sentence was mixed with noise spectrally matched to the long-term average spectrum of the HINT sentences at a Signal to Noise Ratio (SNR) selected from levels expected to be challenging but not impossible for listeners to perceive at least parts of sentences correctly. The actual SNR on each trial was varied randomly per a preset schedule intended to sample the desired range of SNRs uniformly.

After each stimulus presentation, listeners repeated what they heard. This response was digitally recorded via a head-mounted microphone (AKG C-410) at a 16 kHz sampling rate and saved to a file. Subsequently, each recorded utterance was scored as correct or incorrect and transcribed by a human scorer at the word level for use in training the UV engine. This collection of 25 responses to each of the 80 HINT sentences (2000 utterances in all) is referred to as the seed corpus.

2.1.4. rASR and UV Training Procedures

There were three components to the training procedures: (a) training continuous HMM acoustic phone models; (b) designing language models (LMs) that capture the range of possible utterances expected in response to each HINT sentence, and (c) establishing a CM to assess confidence in utterances recognized as CRRESPONSEs. Each of these is discussed briefly below.

2.1.4.1 Acoustic Models

Initial monophone HMMs were trained using HTK on data from the TIMIT database using standard acoustic features (energy plus 12 Mel-frequency Cepstral Coefficients plus their first and second time derivatives calculated from 25 msec Hamming windowed data advanced in 10-msec steps). These single-Gaussian monophone HMMs were force-aligned to the seed corpus and the latter was used to train triphone HMMs with triple-Gaussian Mixture Models. Thus, the final acoustic models were tuned to the triphone contexts, speaker, and acoustic channel characteristics of the 2000 sentences in the seed corpus.

2.1.4.2 Language Models

The language model consisted of a dictionary and set of grammars. The dictionary specified the acceptable pronunciation of each word in the grammars, and was the same for all 80 HINT sentences. The grammar determined the possible word orders and was different for each HINT sentence. It described all of the allowed CRRESPONSEs as well as the entire set of NORESPONSE and ERRESPONSE utterances observed for that HINT sentence in the seed corpus.

2.1.4.3 Confidence measure (CM)

When the rASR engine returned a CRRESPONSE as the recognized item, an additional CM was used to screen for the possibility that an out-of-domain utterance was incorrectly recognized as a CRRESPONSE. A variety of confidence measures have been proposed both for use in general speech recognition and specifically for utterance verification [13]. The approach used here was to train generic consonant (C) and vowel (V) HMMs as ‘background’ models. These were trained as single-Gaussian monophone HMMs using the TIMIT database. A background log likelihood could then be obtained for any utterance by allowing the ASR engine to align the best fitting arbitrary sequence of generic C/V models to the utterance. The difference between this background log likelihood and the log likelihood obtained from the rASR engine was then used as the confidence measure. If the CM exceeded a threshold value, the rASR output was accepted as correct.

To estimate the best threshold value for the CM, all cases from the seed corpus for which the rASR engine returned an element in the CRRESPONSE set were examined and the threshold that corrected the largest number of errors in these data was selected. Based on the seed corpus, the threshold was set to 5.0. Overall, leave-one-out cross-validation (withholding the test case from the training set and retraining) suggested that the prototype UV system would correctly classify about 91.15% of the utterances in the seed corpus without use of the CM. This rose to 93.00% when the CM was used.
3. Evaluation Study

Following training of the UV engine, an evaluation study was conducted to assess its performance in a real-world setting where it was actually used to control the SNR values for standard HINT testing. The study was conducted in a series of blocks with updates to the parameters of the UV engine following each block. This allows reporting of both overall performance, and changes in performance as a function of model updates following each block.

3.1. METHODS

3.1.1. Subjects

The subjects were 25 normal-hearing listeners recruited from the area surrounding the House Ear Institute. None of these listeners participated in the collection of the initial seed corpus. AE was their first and primary language and no attempt was made to restrict or selectively sample AE dialects.

3.1.2. Stimuli

The evaluation experiment used 8 lists of HINT sentences consisting of the four lists for which the UV engine had been trained, and an additional 4 lists for manual administration.

3.1.3. Apparatus

A special version of the HINT software was developed that could be used either to administer the test in the standard fashion (human scoring), or by calling the UV engine to score each response. All listener responses were recorded to disk along with the scoring response of the human or UV engine and saved for later analysis.

3.1.4. Procedure

Each subject was tested separately, seated in a sound dampened chamber wearing headphones (TDH-39P) for stimulus presentation and a head mounted microphone (AKG B-210) for response collection. The order of scoring conditions (Human versus Automatic) was counterbalanced over listeners, and within scoring conditions the order of HINT list presentations and the order of sentences within lists was randomized. The standard HINT protocol was used (e.g., Soli and Wong, 2009), i.e., subjects were instructed to repeat each sentence word for word and to guess if they were uncertain.

The subjects were tested in blocks of five. For the first block the UV engine did not incorporate the confidence measure. Following completion of each block, recorded responses to the four HINT lists used for Automatic scoring were transcribed and used to update the parameters of the UV engine. Thus, each block of five subjects used a slightly different UV engine that incorporated information gained from all previous subjects.

3.2. RESULTS

Overall, threshold SNRs obtained by human and automatic scoring methods differs by just 0.775 dB, or roughly half the standard error of measurement when scoring is done by a human observer [3]. Since the UV engine was updated between successive blocks of 5 listeners, it is also important to consider performance on a block-by-block basis. These results were examined using a mixed model ANOVA with Block (1 - 5) as a grouping factor and Method (human v.s. automatic) as a within-subjects repeated measure. In this analysis, the main effect of Block was non-significant (F[4,20]=1.35, p=0.286), the main effect of Method was non-significant but marginal (F[1,20]=3.88, p=0.063), and the interaction of Block and Method non-significant (F[4,20]=0.57, p=0.69).

Figure 2 shows the mean SNR threshold for each block for Human (circles) and Automatic (triangles) scoring. The figure reveals that a substantial amount of the overall difference between automatic and manual scoring is attributable to differences in blocks 3 and 4. Examining these differences still more closely, revealed that a single listener in Block 3, and two listeners in Block 4 accounted for virtually all of the observed differences in scoring methods. These listeners had thresholds that were 7.75, 4.39, and 3.3 dB higher on average in the Automatic scoring condition, as compared to the Manual scoring condition. A post-hoc analysis of individual listener responses from the evaluation study was undertaken to determine why these three listeners differed from the other 22 listeners, and to identify how to improve the UV engine to avoid such differences.

![Figure 2 Mean HINT threshold by block and method: automatic (triangles); or manual (circles).](image)

4. Discussion

In an adaptive procedure such as the HINT, test-retest variability, that is, the variation in threshold estimates obtained from the same listener over multiple evaluations, is attributable to several sources of error: small differences among the HINT lists; trial-to-trial variations in the listener’s attention; and the accuracy of the tester in scoring the subject’s response. Only the accuracy of the scoring is pertinent to this work. Fortunately, an important advantage of adaptive threshold tests like the HINT is that they are highly robust to small numbers of scoring errors, as long as these errors are not biased.

The fact that the three outliers all obtained Automatic thresholds substantially higher than their Human thresholds suggests that the UV scoring was biased in the direction of rejecting correct responses. Such a bias would drive the adaptive procedure toward higher SNRs. This was verified by analyzing confusion matrices. Overall, 14.2% of the correct responses were misclassified as Wrong and 12.2% of the incorrect responses were classified as Right. However, for the three outliers, more than 40% of the correct responses were misclassified as Wrong, while only 10.5% of their incorrect responses were misclassified as Right.

Examination of the detailed UV engine output revealed that the rASR engine was performing correctly for 2 of these 3 subjects, classifying utterances as Right or Wrong with good
accuracy, but the utterances were being rejected by the CM. The third subject was having utterances incorrectly rejected by the rASR stage of processing. The latter problem is one that suggests simply increasing the amount of acoustic training data will lead to improvements. However, the failure of the CM with the other two subjects suggested that improvements were needed to that component of the overall UV engine.

5. Improving the CM module.

As previously noted, the CM used in the live evaluation experiment was based on a comparison between the alignment of generic C/V models and the utterance as classified by the rASR module. If the overall likelihood of the specific models was not sufficiently better than the likelihood of the generic models, the UV engine would return a classification of **Wrong**. This approach has been extended in two ways. First, in addition to the generic C/V models, we applied an all-phones recognition stage constrained by a bigram LM. The phone bigram probabilities were estimated from a transcription of the Brown Corpus\[14\]. Second, parameters derived from both the all-phones recognition and the C/V alignment were combined with parameters from both the best fitting correct and error models and used to train a logistic regression model that in turn is used to predict the final system response given the initial rASR output and CM measures. Results of applying this revised CM to the data from the evaluation study (in terms of a proportion of correctly classified utterances) are shown in Figure 3. Note that because these were not live data, we cannot estimate thresholds as shown in Figure 2, but can show proportion of correct responses. In Figure 3, the results of the new analysis (triangles) are compared to those from the live sessions (circles). As this figure shows, all blocks have higher proportion of correct calls than the corresponding live blocks. Notably, the large drop in blocks 3 and 4 associated with the questionable subjects have been greatly reduced by the new approach to the CM.

![Figure 3. Proportion of correct response classifications for original (circles) and improved (triangles) CM.](image)

6. Conclusions

A UV engine has been developed and coupled with the HINT system to produce a prototype that operates autonomously to obtain thresholds which, on average, are equivalent to those obtained using standard testing measures. Although the live evaluation study revealed somewhat greater variability for the UV scoring system, and a bias in the type of scoring errors made for 3 of the 25 subjects, the post-hoc data analysis identified the source of these differences and demonstrated that an update to the UV system significantly reduced them. Thus, the fundamental feasibility of the approach has been demonstrated, as well as the ability to adapt and fine-tune the system to improve its performance.

7. Acknowledgements

This work was supported by grant 1R43DC008212 from NIH/NIDCD to the 4th author.

8. References