A Framework for the Evaluation of Microscopic Intelligibility Models

Ricard Marxer[1], *Martin Cooke*[2], *Jon Barker*[1]

[1] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{r.marxer, j.barker}@sheffield.ac.uk
[2] Ikerbasque, Spain
m.cooke@ikerbasque.org

Abstract

Traditional intelligibility models are concerned with predicting the average number of words heard correctly in given noise conditions and can be readily tested by comparison with listener data. In contrast, recent 'microscopic' intelligibility models, which attempt to make precise predictions about a listener's perception or misperception of specific utterances, have less well-defined goals and are hard to compare. This paper presents a novel evaluation framework that proposes a standardised procedure for microscopic model evaluation. The key problem is how to compare a model prediction made at a sublexical granularity with a set of listeners' lexical responses, especially considering that not all listeners will hear a given noisy target word in the same way. Our approach is to phonetically align the target word to each listener's response and then, for each position in the target word, calculate both the probability of a phonetic error occurring and the probability distribution of possible phonetic substitutions. Models can then be evaluated according to their ability to estimate these probabilities using likelihood as a measure. This approach has been built into a framework for model evaluation and demonstrated using a recently released corpus of consistent speech misperceptions and a set of naive intelligibility models.

Index Terms: objective measures, speech intelligibility prediction, evaluation, speech perception

1. Introduction

Traditional speech intelligibility models have been developed primarily as tools for designing and assessing speech communication systems [1, 2, 3, 4]. They are employed to make predictions about the average number of words that may be heard correctly given the specification of a communication channel, e.g. in terms of average noise level, average reverberation characteristics etc. Such models have an unambiguous goal and they can be readily evaluated by comparing their predictions with real intelligibility scores measured using listening tests.

Recently, a new direction in intelligibility modelling has started to emerge: 'microscopic' models [5]. These models attempt to make fine-grained predictions about listeners' responses to speech. For example, such a model might be expected to predict a confusion matrix, to characterise phoneme mis-identifications, or even to predict a listener's lexical response to a specific noisy speech token. It is hoped that by building models of this type, it will be possible to gain an understanding of the auditory mechanisms underpinning speech perception, i.e. such models will have explanatory power.

In contrast to traditional 'macroscopic' models, microscopic models are lacking in any standardised framework for evaluation. Different models have been compared using different data sets and using narrow closed-set designs that cannot be guaranteed to generalise well. For example, Cooke's 'glimpsing' model [5] was evaluated on the basis of its ability to predict identification rates and confusion matrices for a set of consonants presented in noise. Consonants were presented in a VCV context with a fixed vowel, /a/. The model was extended in [6] and evaluated using a task involving recognition of spoken letter-digit grid references. Jurgens et al. [7, 8] evaluated a model based on principles similar to Cooke's but tested the model using German logatomes (CVCs and VCVs). Phatak et al. [9, 10] proposed a model based on the Articulation Index (AI) [1] to predict the relative intelligibility of consonants in CV contexts. Reigner et al. [11] developed a representation called the AI gram – portraying articulation density in the spectral-temporal domain. The AI gram and signal-to-noise ratio were used to explain confusions of CVs in noise, but not the CVs used by Phatak et al.

Given the lack of standardisation it is hard to know how the performances of existing models compare. It is equally unknown how well existing models generalise beyond the conditions under which they have been tested. As a step towards answering these questions, this paper proposes a novel standardised framework for evaluation. The framework is designed to be compatible with existing microscopic models but also to extend the scope of current evaluations in two key respects. First, the framework allows for models to be evaluated using naturally-spoken multisyllabic words. Second, model predictions are evaluated on a truly token-by-token basis rather than on their ability to estimate recognition rates or confusion statistics averaged across entire listening experiments. Moving beyond closed-set responses and averaged predictions has raised a number of theoretical and technical challenges, the solutions of which are the main contribution of this paper.

The remainder of the paper is structured as follows. Section 2 explains how suitable listener data is collected and prepared. Section 3 describes the predictions of this data that models are required to make. Section 4 describes how these predictions are evaluated. Section 5 illustrates the framework using a set of naive intelligibility models. Finally, Section 6 concludes with some discussion of outstanding issues and future challenges.

2. Data

2.1. Listener data

The evaluation framework we present is designed to be able to employ data from any listening experiments in which listeners have been asked to report words heard in response to noisy stimuli. In this paper we illustrate the framework using data from a large scale confusions corpus reported in [12]. The corpus

Copyright © 2015 ISCA

consists of examples of noise-corrupted words that have been consistently misheard. It was acquired by presenting Spanish listeners with 1-3 syllable Spanish target words presented in a range of speech-based noise backgrounds at a range of SNRs. Each stimulus was presented to 15 or more different listeners who were asked to type the word that they heard. If at least 6 listeners provided an identical but incorrect response then the stimulus was considered to have generated a consistent confusion and was included in the corpus. The full corpus consists of responses to 3235 stimuli.

2.2. Aligning listener responses

In order to understand the listener misperceptions in terms of errors at a sublexical level, each reported word is expressed as a phoneme sequence (i.e. corresponding to its pronunciation) and compared to the phoneme sequence corresponding to the target word. Given that there will not generally be a one-to-one correspondence between the two phoneme sequences, the interpretation of the listener's response (i.e. in terms of phonetic errors) depends on how the sequences are aligned.

We propose to use *edit scripts* to represent the alignment between the target and listener phoneme sequences. An edit script is a sequence of operations performed on an input string of symbols that when applied produces a new sequence of symbols. In our context the edit scripts operate on the pronunciation of the presented word to produce the pronunciation of the reported word. We limit the set of possible operations to the following:

- **match** of a phoneme (no confusion)
- **substitution** of a phoneme
- **deletion** of a phoneme
- **insertion** of a phoneme or sequence of phonemes

Each operation is applied to a particular position in the presented word's phoneme sequence. Note, matches, substitutions and deletions are applied to phonemes, whereas insertions occur at positions between phonemes. In the following sections, these positions are indexed as illustrated in Table 1.

Word, γ	maestro														
Phonemes, ϕ^j		m		a		ˈe		s		t		ɾ		o	
Index, j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Table 1: Example of indexing.

We define an edit cost for individual substitution, deletion and insertion operations. We then assume the alignment between the listener's response and the presented word is the one that minimises the total edit cost. We have currently chosen a non-informative set of costs for matching (0), substitution (1), deletion (1), insertion (1×inserted phoneme), however the evaluation could be conducted with different sets of costs.

Note, it is possible that several alignments may share the same minimum cost – e.g. see Table 2. Even though these alignments have the same cost, they will align phonemes differently and imply a different set of phonetic confusions has occurred. In these cases the alignment is selected which gives the model being evaluated the best score (i.e. 'benefit of the doubt').

3. Tasks

We propose to evaluate intelligibility models according to their ability to perform three separate tasks relating to listener response predictions of increasing detail and difficulty: *confusion frequency prediction*, *confusion characterization prediction* and *full confusion prediction*. In the first task, models need only predict the probability of confusions occurring at each phoneme position. In the second, models must predict what each phoneme will be confused with (represented as a probability distribution over the possible substitutions). Finally, the last task is to predict the complete word responses heard by the listeners - again, represented as a probability distribution over possible responses.

	maestro														
		m		a		ˈe		s		t		ɾ		o	
ministro		m	i	n		ˈi		s		t		ɾ		o	
		m		i	n	ˈi		s		t		ɾ		o	
		m		i		n	ˈi	s		t		ɾ		o	

Table 2: Example taken from the corpus of multiple alignments for the same pair of pronunciations.

For all tasks it is assumed that the model can have access to both the target utterance audio and to the masker audio (i.e. the audio signals that were mixed to construct the noisy stimulus). The evaluation framework should also specify a common lexicon, phoneme set and pronunciation dictionary, so that models can be more readily compared. These should be selected appropriately for the listening data being employed.

The three tasks are described in full detail below.

3.1. Confusion Frequency Prediction

The goal of the confusion frequency task is to predict how often confusions occur in different parts of the presented utterance. In preparation, we first align the pronunciation of the reported words to the pronunciation of the presented word using edit scripts, as presented in Section 2.2. This representation allows us to isolate the confusions occurring at each phoneme location.

For a given token, let k_j be the number of aligned responses from listeners for which the phoneme at index j is different from the one presented (confusion), and let n be the total number of responses from listeners. For example, Table 3 shows the number of confusions k_j for a token in the dataset with $n = 16$. We assume the number of confusions to follow a binomial distribution $k_j \sim Binomial(n, p_j)$ with the parameter p_j being the probability of encountering a confusion at index j for a given response.

Under this assumption the task of a micro-intelligibility model is to provide a probability of confusion, p_j, for each of the positions j in the pronunciation of the presented word. For simplicity, in the current framework, these probabilities are considered to be independent.

3.2. Confusion Characterization Prediction

The objective of the confusion characterization task is to provide a more detailed description of the confusions. As in the previous task we regard the confusions as edit scripts. The task is to predict what each phoneme in the presented word will be heard as. For each presented phoneme, this prediction will be a set of phonemes or phonemes sequences and their respective probabilities.

For this task the random variable defined for each presented phoneme index j is categorical. It can have any of the following values: *match* (i.e. no confusion), *empty* (i.e. deletion), any possible phoneme label, or any possible sequence of phonemes (i.e. multiple phonemes can be inserted at a position). The

		Presented Word														
		maestro														
	Index, j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Responses	**Count**		m		a		'e		s		t		ɾ		o	
nuestro	6		n		w		'e		s		t		ɾ		o	
diestro	3		d		j		'e		s		t		ɾ		o	
maestro	3		m		a		'e		s		t		ɾ		o	
ministro	3		m	i	n		'i		s		t		ɾ		o	
siniestro	1	si	n		j		'e		s		t		ɾ		o	
Confusions, k_j		1	10	3	13	0	3	0	0	0	0	0	0	0	0	0

Table 3: Example taken from the corpus of the confusion frequency task.

probability of these outcomes is described by a multinomial distribution. (In order that there is a finite set of outcomes, only insertions of phoneme sequences of up to a fixed length will be considered to be possible. See Section 5.2 for further discussion.)

Similarly to the previous task we let k_{j,ϕ_i} be the number of responses for which the phoneme ϕ^j at the presented index j is aligned with the phoneme or sequence of phonemes ϕ_i in the response. As before n denotes the total number of responses for the token. We assume the confusions to follow a multinomial distribution $(k_{j,\phi_1}...k_{j,\phi_i}...k_{j,\phi_I}) \sim Multinomial(n, (p_{j,\phi_1}...p_{j,\phi_i}...p_{j,\phi_I}))$ with parameters p_{j,ϕ_i} being the probability of attributing to index j the phoneme or sequence of phonemes ϕ_i. I being the number of different phonemes or sequences of phonemes that appeared for index j of the presented word among all the responses.

3.3. Full Word Confusion Prediction

In this last task the goal is to predict the words reported by the listeners. To deal with homophones, the predicted word and a reported word are considered to match if and only if they have the same pronunciation, i.e. identical phoneme sequences. Again, because there are multiple listener responses, the model is expected to estimate the distribution of the reported words, i.e. accompanying the list of predicted words with a corresponding set of probabilities.

For this case we let k_{γ_i} be the number of responses whose pronunciation is γ_i. Again n is the number of responses for the token. As in the previous task we assume the responses of the listeners to follow a multinomial distribution $(k_{\gamma_1}...k_{\gamma_i}...k_{\gamma_I}) \sim Multinomial(n, (p_{\gamma_1}...p_{\gamma_i}...p_{\gamma_I}))$ with parameters p_{γ_i} being the probabilities of response of a word with pronunciation γ_i.

4. Evaluation Procedure

In the previous section we described how models are asked to express their beliefs about listening errors in terms of probability distributions. This formalism allows a clear and objective route for evaluation. We simply rank models according to the probability with which their suggested distributions would generate the observed data, i.e. the best model is the one whose parameters have the maximum likelihood.

Formally, the measure of quality that we propose for the tasks are the likelihoods of the data k_j, k_{j,ϕ_i} and k_{γ_i} given the probabilities p_j, p_{j,ϕ_i} and p_{γ_i} predicted by the intelligibility models. These likelihoods are the probability mass functions of the binomial and multinomial distributions. For practical reasons we use the log-likelihoods, leading to the following scores:

$$\mathcal{L}_j^{freq} = \log\left(\binom{n}{k_j} p_j^{k_j}(1-p_j)^{n-k_j}\right) \quad (1)$$

$$\mathcal{L}_j^{char} = \log\left(\frac{n!}{k_{j,\phi_1}!...k_{j,\phi_I}!}{p_{j,\phi_1}}^{k_{j,\phi_1}}...{p_{j,\phi_I}}^{k_{j,\phi_I}}\right) \quad (2)$$

$$\mathcal{L}^{full} = \log\left(\frac{n!}{k_{\gamma_1}!...k_{\gamma_I}!}{p_{\gamma_1}}^{k_{\gamma_1}}...{p_{\gamma_I}}^{k_{\gamma_I}}\right) \quad (3)$$

The likelihoods of the confusion frequency and confusion characterization tasks are defined for each index j of the presented words. In order to obtain a score for a token we average over all of the token indices so that $\mathcal{L}^{freq} = \frac{1}{J}\sum_j^J \mathcal{L}_j^{freq}$ and $\mathcal{L}^{char} = \frac{1}{J}\sum_j^J \mathcal{L}_j^{char}$, where J is the number of indices in the given token. Finally we average the token scores across the corpus to obtain: $\overline{\mathcal{L}^{freq}}$, $\overline{\mathcal{L}^{char}}$ and $\overline{\mathcal{L}^{full}}$.

We note that a particular response could have multiple pronunciations in the lexicon and even a single pronunciation can be aligned in multiple (equally good) ways with the pronunciation of the presented word. This raises a difficulty because the different pronunciations and alignments will produce different values of k_j, k_{j,ϕ_i} and k_{γ_i}. Which is the correct one to use?

In the proposed evaluation method, the use of lexicons and the phonetic alignment are tools used to estimate how phonemes in the target word have been transformed. The ground-truth of how the confusion are produced or what phonemes are correctly perceived is not available from the experimental data, and when alignments have equal edit distances there is no way to know which is more likely to be correct. Therefore, to avoid unfairly penalising a model by making an incorrect guess, we propose applying the benefit of the doubt and consider all possible pronunciations/alignments of all the different responses and only keep the combination leading to the highest score for each token.

5. Reference Models

We complement the framework with three reference 'pseudo'-models. An *Oracle* model that uses knowledge of the listener data to make perfect predictions; a *statistical* model that uses macroscopic statistics on the corpus; and a *random* model that makes non-informative predictions. These systems provide benchmark scores against which users can gauge true models, e.g. a model adding real 'microscopic' value should perform better than the random and macroscopic statistical models.

5.1. Oracle Model

The Oracle model makes the optimal predictions in each of the tasks using the actual experimental ground truth data. This model serves as an upper bound for the evaluation measures and can be used to provide scores relative to the maximum attainable on a per-phoneme or per-token basis.

Given the probabilistic distributions assumed in the tasks, the probabilities leading to the highest likelihood will be:

$$p_j^o = k_j/n \qquad \forall j \quad (4)$$

$$p_{j,\phi_i}^o = k_{j,\phi_i}/n \qquad \forall j,\ \forall i \quad (5)$$

$$p_{\gamma_i}^o = k_{\gamma_i}/n \qquad \forall i \quad (6)$$

5.2. Random Model

The random model is the result of applying a uniform distribution among all possible choices in each of the tasks. We may consider that the model does not make use of any prior information.

If $I[m]$ is the number of phoneme sequences of length m that may be inserted between phoneme positions and L is the maximum length of an inserted sequence; W is the number of pronunciations in the lexicon, then the predicted probabilities of such a model are:

$$p^r_j = 0.5 \qquad \forall j \quad (7)$$

$$p^r_{j,\phi_i} = \begin{cases} 1/(\sum_{m=1}^{L} I[m]+1) & \text{if } j \in \mathbb{E} \text{ and } \forall i \\ 1/(I[1]+1) & \text{if } j \in \mathbb{O} \text{ and } \forall i \end{cases} \quad (8)$$

$$p^r_{\gamma_i} = 1/W \qquad \forall i \quad (9)$$

In the first task, an equal probability between observing a confusion and not leads to a probability of 0.5. In the second task, the possible outcomes of a confusion depends on whether phonemes (odd index values) or spaces between phonemes (even index values) are being considered. At phoneme positions, there are $I[1]$ outcomes (i.e. all individual phonemes) plus the deletion ($\varnothing$). When considering the spaces between phonemes, the outcomes are all possible phoneme sequence insertions of length 1 or more, plus the non insertion ($\varnothing$). The last task choices are all distinct pronunciations (W).

5.3. Statistical Model

The statistical model is based on the macroscopic statistics of the dataset. The probabilities predicted by this model are the same for all tokens. The predictions are only made based on the total number of each of the edit operations (Section 2.2) that occurred in the dataset.

To formulate the model the following counts are made. k^m, k^r and k^d are the counts of matches, replacements and deletions at phoneme positions. $k^i[m]$ for $m \in [0..L]$ are the counts of insertions of m-length phoneme sequences, $k^i[0]$ being the non-insertions. k^c and k^o are the counts of responses that match and do not match the presented word respectively. $k^{pa} = k^m + k^r + k^d$ and $k^{ia} = \sum_m^L k^i[m]$ being the total counts of operations at phoneme and inter-phoneme positions respectively, and $k^{wa} = k^c + k^o$ being the total counts of listeners responses. l_{ϕ_i} is the amount of phonemes in phoneme sequence ϕ_i.

Given the above counts, the model probabilities are defined,

$$p^b_j = \begin{cases} \sum_{m=1}^{L} k^i[m]/k^{ia} & \text{if } j \in \mathbb{E} \\ (k^r + k^d)/k^{pa} & \text{if } j \in \mathbb{O} \end{cases} \quad (10)$$

$$p^b_{j,\phi_i} = \begin{cases} k^m/k^{pa} & \text{if } j \in \mathbb{E} \text{ and } \phi_i = \phi^j \\ k^d/k^{pa} & \text{if } j \in \mathbb{E} \text{ and } \phi_i = \varnothing \\ k^r/(k^{pa} \cdot (I[1]-1)) & \text{if } j \in \mathbb{E} \text{ and } \phi_i \neq \phi^j \\ k^i[0]/k^{pa} & \text{if } j \in \mathbb{O} \text{ and } \phi_i = \varnothing \\ k^i[l_{\phi_i}]/(k^{pa} \cdot I[l_{\phi_i}]) & \text{if } j \in \mathbb{O} \text{ and } \phi_i \neq \varnothing \end{cases} \quad (11)$$

$$p^b_{\gamma_i} = \begin{cases} k^c/k^{wa} & \text{if } \gamma_i = \gamma \\ k^o/(k^{wa} \cdot (W-1)) & \text{if } \gamma_i \neq \gamma \end{cases} \quad (12)$$

where ϕ^j is the phoneme at the j^{th} position of the presented word γ. $\varnothing$ represents a null phoneme (deletion or non-insertion).

6. Discussion

We present the scores of the proposed models in Table 4. These results serve as a comparable reference for future microscopic intelligibility models: all models should perform between the bounds of chance (*random*) and ideal predictions (*oracle*). Good microscopic models should hope to outperform the model using macroscopic statistics (*statistical*).

model	$\overline{\mathcal{L}^{freq}}$	$\overline{\mathcal{L}^{char}}$	$\overline{\mathcal{L}^{full}}$
oracle	-0.645	-1.122	-4.689
statistical	-3.747	-10.952	-172.463
random	-7.631	-128.821	-184.831

Table 4: Results of the measures on the models.

The tasks vary greatly in difficulty. Note, the range of the scores for the first task ($\overline{\mathcal{L}^{freq}}$) is considerably lower than the others. This reflects the smaller solution space of this task, which consists in predicting a parameter of several Binomial distributions. This is also indicative of the greater difficulty of the second and third tasks where a model must predict multiple parameters of a Multinomial distribution. In all tasks there is significant space for improvement between the baseline macroscopic model (*statistical*) and the optimal microscopic model (*oracle*).

The *random* model presents a very low score in the phoneme characterisation task ($\overline{\mathcal{L}^{char}}$) compared to the *statistical* model. This large difference is mainly due to the fact that the *statistical* model takes into consideration the length of the inserted phoneme sequence in the predictions (e.g. single phonemes are more likely to be inserted than phoneme sequences of length 5). In contrast, the *random* model which assumes a uniform distribution among all possible insertions of all lengths, which leads to a very low predicted probability for each possible insertion.

The low scores of both the *statistical* and *random* models in comparison to the *oracle* in the last task ($\overline{\mathcal{L}^{full}}$) illustrates the main problem we are trying to solve with the evaluation framework. Correctly predicting the distribution of reported words given an open vocabulary is a difficult task, which justifies the proposal of the two other tasks, when evaluating microscopic intelligibility models.

7. Conclusions and Dissemination

We have presented a framework to evaluate and compare microscopic intelligibility models. The proposed framework consists of i/ a set of tasks for the models to solve, ii/ a set of measures to assess the accuracy of the performance and iii/ three reference pseudo-models. The reference models provide scores by which future microscopic models can be gauged.

Several future challenges remain open, such as the way in which to perform more detailed analysis than average summaries of the scores, and a study of how the choice of evaluation parameters (phone sets, edit costs, lexicon) may affect the results. However this proposal already opens the door to a more consistent and rigorous comparison and analysis of microscopic intelligibility modelling.

The framework has been fully implemented and has been made publicly available in the form of an online evaluation service. For full details of how to access the evaluation data and for tutorials explaining how to prepare submissions, readers are directed to `http://goo.gl/gFFw0H`.

8. Acknowledgements

This work was funded by the INSPIRE-ITN (European Commission FP7-PEOPLE-2011-290000) project.

9. References

[1] ANSI, *Methods for the Calculation of the Articulation Index*, American National Standards Institute, 1430 Broadway, New York, NY 10018, USA, 1969.

[2] ——, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, 1430 Broadway, New York, NY 10018, USA, 1997.

[3] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90–119, 1947.

[4] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, pp. 318–326, 1980.

[5] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[6] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication.*, vol. 49, pp. 402–417, 2007.

[7] T. Jurgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2635–2648, 2009.

[8] T. Jürgens, T. Brand, and B. Kollmeier, "Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model," in *Proc. Interspeech*, 2007, pp. 410–413.

[9] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2312–2326, 2007.

[10] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1220–1233, 2008.

[11] M. S. Regnier and J. B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2801–2814, 2008.

[12] M. A. Tóth, M. L. García Lecumberri, Y. Tang, and M. Cooke, "A corpus of noise-induced word misperceptions for spanish," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. EL184–EL189, 2015.