Who is getting interrupted?
Generalisability in End of Turn Prediction

B152029

Wordcount:7996

Master of Science

Speech and Language Processing

School of Philosophy, Psychology & Language Sciences

University of Edinburgh

2020

Abstract

Generalisability can be viewed in multiple senses, and has a complex relationship with model adaptation. I explore how models of turn-taking generalise, with long-term motivation of enabling less exclusionary models. I train general turn-taking models on Maptask, Switchboard and a Combined corpus. Acoustic-only models perform better than linguistic-only; bi-modal models perform best. In cross-corpus tests, performance is generally best training and testing on the same corpus, but there is variation for linguistic-only models, perhaps due to over-fitting. Models trained on Combined do not out-perform other models despite more data, suggesting they are not generalising across corpora. On Switchboard, acoustic features generalise better than linguistic features; this is not clear on Maptask.

I explore intra-corpus generalisation, looking at role differences. Maptask has a role difference, but performance is best on models trained on both roles. Switchboard has no role difference, and little variation when models are trained on individual roles. There are considerable differences in model performance on individuals in both corpora, especially Maptask, possibly due to differences in formality and rapport, or for under-represented groups. I explore the idea of adapting metric parameters, experimenting on Prediction at Onset, to enable "quick win" model adaptation without re-training neural network parameters.

Acknowledgements

Thank you to Catherine Lai, for her wise, knowledgeable and – above all – understanding and human supervision in the face of a global pandemic.

Thank you to Simon King for his practical and kind advice in these circumstances.

Thank you to Lena and Elliot for your fun and thoughtful collaboration, and for reaching out at times when it really meant a lot.

Thank you to Brooke, Emil, Karolina and the rest of the SLP cohort for staying in touch from afar, and always being happy to talk things through.

Thank you to everyone who brought me food when I was stuck inside, particularly those who brought plenty of chocolate.

Thank you to Carine and Alasdair, Rick and Julia for your unwavering support when things got particularly tough.

Thank you to Sharon and Rebecca for encouraging me to take the leap, and to Ellie, Mim and Rose for always cheering me on.

And Andrew, thank you for always thinking of me, and for coming on this adventure too.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*B152029*)

Table of Contents

1. Introduction

"I am not a number." – *Number Six, The Prisoner*

When humans converse, they predict when the other person will finish speaking, so they can plan their response. This allows the speakers to communicate efficiently, minimising both gaps and overlap (Sacks et al., 1974). For conversational systems to have efficient and natural-seeming conversations with humans, they must be able to mimic this ability. The simplest method is to use a pause threshold to decide whether a speaker has finished speaking, but this does not allow conversation that is natural or efficient, because it is not predictive (Skantze, 2017). More complex approaches provide a more natural experience, by modelling when a speaker is likely to stop speaking, or classifying silences as pauses or turn-ends.

As with many neural networks trained on speech and language data, models for End of Turn Prediction can perform well on data similar to what they have been trained on, but we would like them to generalise, be that to other tasks, recording conditions, speakers, speaker-demographics, corpora or languages. There will always be some bias in any data set, or some types of speakers represented more than others. It is therefore important to understand whether and how models for End of Turn prediction can generalise. When they cannot, we must understand who is being excluded. This is a matter of fairness: conversation systems should listen and respond to speakers from different backgrounds, completing different tasks.

When End of Turn Prediction is not working effectively, speakers may experience an unnatural pause before getting a response, or be interrupted mid-utterance. Without considering who gets interrupted, or which tasks are more difficult to accomplish, we risk perpetuating existing social inequalities; our systems may consistently interrupt those who are already regularly interrupted in human conversation. Understanding this is a step towards solving problems, thereby enabling End of Turn Prediction modules and, by extension, dialogue systems that serve all users.

1.1 What is Generalisability?

Generalisability can be interpreted in various ways. Firstly, in a mechanical sense: can a model trained on one data-set be used to test on another? However well a model can predict turn-taking, it may not work on data that is not in the correct format, or does not have any required annotations. These architectural considerations should be kept in mind, particularly when selecting features.

Secondly, in terms of model performance: Aggarwal, 2018 describes *generalisability* as "The ability of a learner to provide useful predictions for instances it has not seen before"; essentially the opposite of over-fitting. Is the model learning truly predictive insights about the data, or just memorising patterns inherent to this data-set? Since neural networks are black boxes, it is difficult to know what the model is learning, even if we had a clear idea of what cues we wanted models to learn.

Thirdly, *generalisability* can describe how well a model performs on another corpus. Has it only learned patterns relevant to the data in the corpus it was trained on, or has it learned more general rules that can apply to other corpora?

These definitions are all inter-twined; we must bear them all in mind. Although there is no single, clear definition, my motivation in considering *generalisability* is to look beyond chasing metrics as a way of judging model success, and understand how models are really performing for underrepresented groups and individuals.

As Futoma et al. note about machine learning in a health-care setting: models may or may not be able to generalise from one context to another. Although it is desirable if they do, lack of generalisability does not mean we should discard a model that is effective in one setting. However, we should seek to understand how and when models are generalising so we know if a model will be useful in a specific context, and which users it is serving well. Finally, if we find models are unable to generalise well for turn-taking, we could explore ways to adapt models to new contexts, as an alternative solution to the generalisability problem.

2. Related Work

2.1 Human Turn-taking

Since conversation is central to language use, Sacks et al., 1974 hypothesise that other aspects of linguistic structure are designed to help turn-taking. Humans use a range of sophisticated cues to guide their turn taking behaviour, many of which are linguistic. Cues include: filled pauses, incomplete syntax, flat pitch, and lack of eye contact to indicate holding the floor; rising or falling pitch, and shifting gaze towards the addressee to indicate yielding (Skantze, 2017). Information density is another: Dethlefs et al., 2016 show humans prefer overlaps in speech when information density is low. The more turn-release cues a speaker uses, the more likely another speaker will take the floor (Duncan, 1972).

Sacks et al., 1974 also propose that humans aim to minimise gaps and overlaps between turns. This is not universally accepted: Heeman and Lunsford, 2017 suggest the goal is to optimise task completion and efficiency; speakers contribute to the conversation if they have something relevant to say.

2.1.1 Prediction Ceiling

Sacks et al., 1974 assert that whether or not a silence constitutes a turn-yield or a pause is "transformable". Turn transitions are negotiated between speakers (Heeman and Lunsford, 2017); there may not be a definitive answer as to whether a given point in conversation is appropriate for a speaker to take the floor. Therefore, a perfect performance by an end-of-turn prediction system is not likely to be possible, though I am not aware of work on what the ceiling may be.

2.2 Existing Approaches to End-of-Turn Recognition

2.2.1 Thresholding

Thresholding is the traditional approach to turn–taking in dialogue systems; the system waits for silence of a certain length, and takes this to mean the user has finished speaking (Skantze, 2017).

This is problematic, since "pauses between turns are sometimes shorter than pauses within turns" (de Kok and Heylen, 2009). If a user makes a long pause before they finish speaking - which is common under a high cognitive load - the system will interrupt them. This can be avoided by increasing threshold duration, but then increases system latency when the user has finished speaking (Arsikere et al., 2015).

2.2.2 IPUs

A more nuanced approach uses Inter-Pausal Units (IPUs) - periods of speech from a single speaker, with no silence longer than a threshold. A turn is made up of one or more IPUs from the same speaker (Skantze, 2016). After each IPU, a model predicts whether the speaker is pausing or has finished speaking (Skantze, 2017). Whilst this approach reacts to periods of silence, humans are proactive (Tice and Henetz, 2011); we predict when the next turn-transition will happen, enabling us to pre-plan responses and quickly take the floor.

2.2.3 Continuous Prediction

For turn-taking systems to operate predictively, they must operate incrementally. They must start making predictions given a minimal amount of input, and continue to do so at every time step, rather than waiting for IPU boundaries (Skantze and Schlangen, 2009). Roddy et al., 2018b implements this by predicting the next 60 frames of voice activity for a speaker at every frame, and I follow that.

2.2.4 General Models

A system must model many phenomena to mimic natural turn-taking behaviour: should the model respond when there is a brief silence, or will the user continue; where are back-channels appropriate; is the user starting a long or short utterance; will the user interpret a pause as a cue the system is yielding the floor (Skantze, 2017)? Skantze, 2017 and Roddy et al., 2018b implement models that make *general* predictions, which I follow. The model outputs probabilities of future voice activity, which are used to predict other turn-taking phenomena, as outlined in Section 4.4.

2.2.5 Multi-Task Learning

Aldeneh et al., 2018 train a model to predict dialogue act as a secondary task, encouraging the model to use information about speaker intentions as part of its turn-taking predictions. They find the model gives a slight improvement over their baseline, but prediction success varies considerably by dialogue act. These benefits would be challenging to bring to a live system, as dialogue act annotations may not be available. However, use of multi-task learning to improve performance of turn-taking models in the context of cross-domain turn-taking models is scope for future work.

2.3 Generalisability in Turn-Taking

2.3.1 Cross-Domain Turn-Taking

Ward et al., 2019 find large differences in model performance across Japanese corpora, for example finding far better performance on telephone speech than on Maptask. They attribute this to task differences, and levels of formality, and highlight the need for adaptation techniques for turn-taking. Heeman and Lunsford, 2017 also found turn-taking behaviour depends on the nature of the dialogue. Yang and Heeman, 2010 found turn conflicts are more likely when there are time constraints, which implies that turn-taking behaviour is related to "the importance of task completion" (Selfridge and Heeman, 2010). Arsikere et al., 2015 investigate the difference between a formulaic template-based corpus, and a more free-flowing one. They find models trained on spontaneous speech perform well on formulaic speech, but the reverse is not true, suggesting that some – but not all – data-sets contains cues that allow models to generalise well.

2.3.2 Cross-Lingual Turn-Taking

The difference between languages can be viewed as an extreme version of the difference between dialects – or other domains. If models can generalise across languages, this suggests they could generalise across less extreme domain differences.

Although Sacks et al., 1974 claim their systematics of turn-taking apply to languages other than English, there is no consensus as to whether there are universals in turn-taking. Stivers et al., 2009 find universals of minimising gaps and overlaps across languages, with a "cultural calibration" regarding acceptable size of gaps and overlaps. Gravano et al., 2016 found listeners could predict turn-taking phenomena better than chance from recordings of a language they did not speak. Conversely, some studies focus on culture-specific ways in which turn-taking systems vary, though with the view that universals may emerge (Bauman and Sherzer, 1989 p.8).

Brusco et al., 2017 train models on English and Spanish. In cross-lingual testing they find performance is lower – but still above chance – when there is a mismatch between languages in the training and test sets. So, for example, a turn-taking module trained on a different language could be used as a starting point for a dialogue system on a low-resource language. Similar to the cross-corpus findings of Arsikere et al., 2015, Brusco et al., 2017 find that performance degradation is not symmetrical across languages. They suggest that if one language has a richer inventory of turn-taking cues, rules from that language could generalise better to the other language than vice versa. Ward et al., 2019 also find commonality across languages, but assume a separate model must be trained on each language.

3. Data

3.1 HCRC Maptask Corpus

Maptask is an unscripted but carefully controlled exercise. Two participants are given maps; the *instruction giver* (G) has a route on theirs, and must explain to the *instruction follower* (F) how to draw that route (Anderson et al., 1991). Focus on the task distracts participants from their speech, resulting in natural, spontaneous and informal data (Anderson et al., 1991). The corpus predominantly represents Scottish English.

I used the same train test split as in Roddy et al., 2018b.

3.2 Switchboard NXT Corpus

Switchboard is a corpus of "spontaneous conversational speech", with conversations in various dialects of American English. Participants are connected to a stranger via telephone and given a discussion prompt (Godfrey et al., 1992). Although one of the speakers (A) initiates the call, the task is unstructured and there are no clear roles.

Following Gruzin, 2020, I use only conversations with timestamps on orthographic words, from Switchboard NXT (Calhoun et al., 2010). Since Switchboard was designed for speaker identification, many speakers appear in multiple conversations (Godfrey et al., 1992). To avoid speakers appearing in both test and training sets, I used all conversations with repeating speakers for training, and remaining conversations for test.

3.3 Combined Corpus

This is an amalgamation of Maptask and Switchboard. The training set comprises the combined training sets from both corpora, and the test set comprises the combined test sets.

3.4 Preprocessing

I used and adapted scripts from Roddy, 2018 and Gruzin, 2020.

3.4.1 Split Channels

Dialogues are split into channels: a single audio file contains the speech of each participant. Each side of the conversation is treated as a separate piece of training data. Some evaluation metrics (see Section 4.4) are calculated on voice activity predictions from one channel only, whilst others make predictions by comparing model outputs for both speakers.

3.4.2 Extract Acoustic Features

Acoustic features are eGeMAPS, extracted using OpenSmile (Eyben et al., 2010). GeMAPS are a minimalistic parameter set designed to increase comparability of results, prevent overfitting and improve cross-corpus generalisation. GeMAPS have been found to produce remarkably comparable results to models using thousands of features. eGeMAPS are an extended version of the base GeMAPS, with additional cepstral descriptors added, to the base "prosodic, excitation, vocal tract, and spectral descriptors found to be most important in previous work", giving 88 total parameters (Eyben et al., 2016).

3.4.3 Extract Linguistic Features

In this paper, linguistic features are words. Roddy et al., 2018a find using parts-of-speech (POS) generally less successful than using words on Maptask data, although POS improve Prediction at Onset. There is potential for future work on whether other, or additional, linguistic features, could help models generalise.

Words are processed into indices, and when two words are found in the same time-step, they are tokenised as a single word.[1] Word embeddings – vectors representing the word – of length 64 are jointly trained with the network.

Roddy found pre-trained GloVe embeddings did not improve performance; he speculates GloVe may provide more benefit on a larger, more conversational corpus such as Switchboard (M. Roddy, personal communication, July 19, 2020). This is scope for future work, as is investigating potential improvements from other pre-trained embeddings.

[1] The intention is to capture multi-token words such as "it's". This is problematic, but an exploration is beyond the scope of this paper.

4. Methodology

4.1 LSTMs

Recurrent neural networks (RNNs) are known to suffer from vanishing gradients (Jurafsky and Martin, 2019). LSTMs help mitigate this, as detailed in Figure 4.1.

4.2 Multi-scale LSTMs

With conventional LSTMs, all data must be input at the same timescale. For example, features could be sampled from the data every 10ms, or every 50ms. For multi-modal models, where modalities naturally have features at different timescales, this is problematic; when using acoustic and linguistic features, should we up-sample the linguistic features, potentially confusing the model, or down-sample the acoustic features, potentially losing valuable data?

Roddy et al., 2018b tackle this with a multi-scale architecture (Figure 4.2). In experiments with linguistic and acoustic features, they find combining modalities using different timescales improves model performance over using single modalities, or combining modalities at a single timescale.

4.3 Hyper-parameters

Roddy et al., 2018b perform a grid search to optimise hyper-parameters for every model configuration. I have not done this, since it is not my goal to maximise performance. I keep the same hyper-parameters for every model configuration, to facilitate comparability across models. I speculate that models with highly tuned hyper-parameters may not generalise as well, although exploration of this is left to future work.

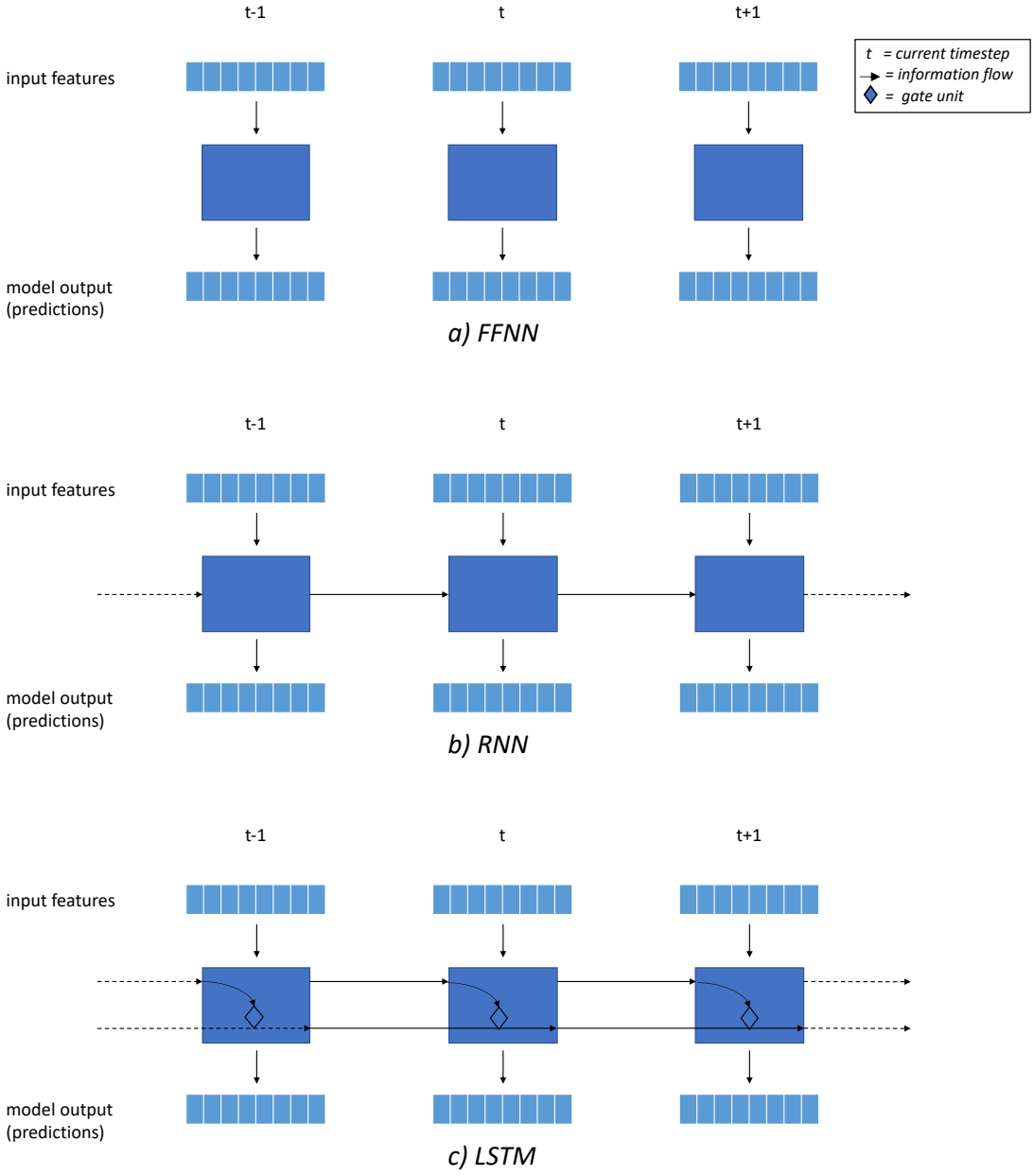

Figure 4.1: Simplified schematics of neural network architectures. In a), there is no flow of information between time-steps. In b) time-steps are connected, but the unrolled network is too deep, causing vanishing gradients (Olah, 2015). In c) time-steps are connected, but gate units manage what information is passed between each time step, mitigating the vanishing gradient problem (Hochreiter and Schmidhuber, 1997).

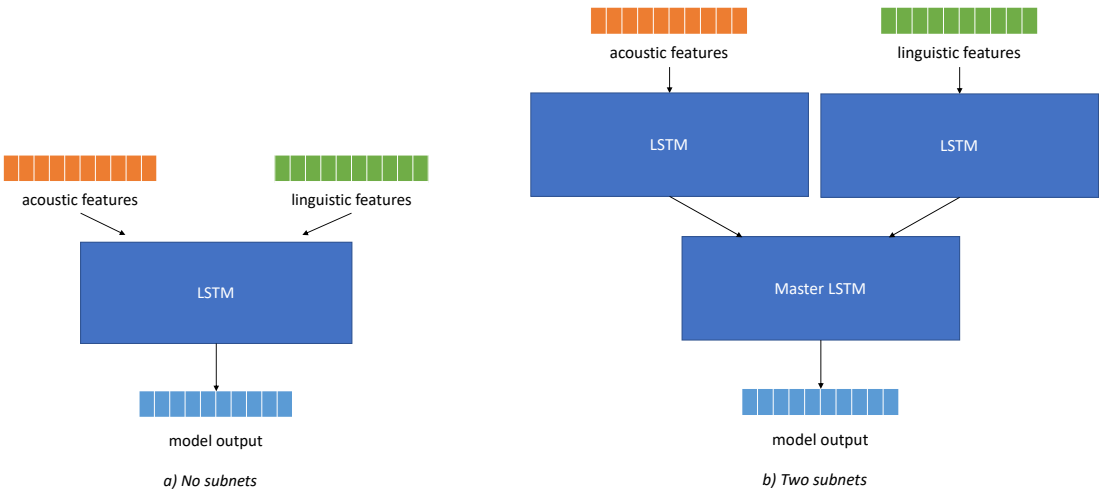

Figure 4.2: Multi-scale LSTM architecture, after Roddy et al., 2018b. In a), there are no subnets; a single LSTM models features from both modalities. In b), there are two subnets; a separate LSTM models each modality, each can operate at a different timescale. These are combined using a master LSTM, which concatenates the current hidden state from each of the LSTM subnets.

I use the hyper-parameters provided in Roddy, 2018: *learning rate* 0.01; *hidden layer size* 60; *patience* 10.

4.4 Evaluation Metrics

4.4.1 BCE Loss

All models are trained to minimize Binary Cross Entropy (BCE) loss, following Roddy et al., 2018a who find it produces better results than Mean Absolute Error. In results, BCE Loss is reported against the test set.

4.4.2 Prediction at Pauses

At every pause of a minimum set length, we use the model outputs to predict who will speak next. Taking the mean of each speaker's voice activity probabilities after the pause; the speaker with the highest mean is predicted to speak next. Following Roddy et al., 2018b, I report F1 score of these predictions after 50ms and after 500ms.

Averaging is a simplistic way of using these model outputs; exploration of improving performance by feeding the outputs into an FFNN can be found in Gruzin, 2020.

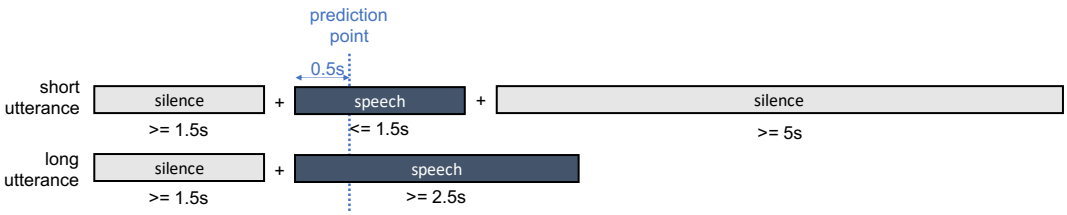

Figure 4.3: Definition of *short* and *long* utterances for Prediction at Onset, following Roddy et al., 2018b. 0.5s from the start of the utterance a prediction is made as to whether it will be *short* or *long*.

4.4.3 Prediction at Onset

This predicts whether an utterance will be *short* or *long* (defined in Figure 4.3). *Short* utterances are similar to back-channels (Roddy et al., 2018a). At the prediction point, the mean of the model output probabilities is compared to a threshold value; if the mean is above the threshold, it will predict a *long*, utterance, otherwise *short*. Threshold value is determined using a R.O.C. curve, maximising correct predictions on the training data.

As with Prediction at Pauses, this is a simplistic way of using model outputs. Performance may also be improved using an FFNN to make classifications using the output predictions.

4.4.4 Average F1

Where all F1 scores described above are reported for a test set on a given model, I also report mean F1 to aid comparison between models. This gives an indication of overall model performance, assuming the three scores are equally important. This assumption should be revisited in light of specific applications.

5. Experiment: Turn-Taking Prediction on Various Corpora

I reproduce results from Roddy et al., 2018b on Maptask and other corpora. I hypothesise that Roddy et al., 2018b's findings about model configurations will hold on all corpora: modelling acoustic and linguistic modalities at different timescales will improve performance, acoustic features will be best modelled at 10ms, and linguistic features at 50ms. There are differences between corpora, especially in terms of dialect and task. However, since the speakers are mutually intelligible, there should be similar types of cues that models can learn from.

I expect results on acoustic-only models to vary less between corpora, since vocabulary is more task specific than acoustics, although I would expect both to vary somewhat by dialect. Roddy et al., 2018a find that acoustic features are generally more useful than linguistic features for end-of-turn prediction, and I believe this is due to better generalisation from acoustic features.

5.1 Method

I re-run experiments on Maptask from Roddy et al., 2018b to reproduce the results. I follow the initial paper, where five models are trained for each no-subnet architecture, and three for each two-subnet architecture, with results averaged over each run of the model. Since they are available, I also include results of models my colleagues trained on Maptask in my averaged results, to increase statistical significance of those scores.

I also train models on Switchboard, and Combined, averaging across 3 and 5 models.

5.2 Results

Figures 5.1 and 5.2 show results of the original Roddy et al., 2018b experiments, alongside my results for each corpus. The original F1 scores are generally higher than my reproduced Maptask results, likely a result of hyper-parameter tuning, which I omitted (see Section 4.3). Overall patterns are similar; correlation between corpora for all metrics is above 90% with $p<0.01$ on both Spearman and Pearson measures.

Average F1 score (Figure 5.1), Roddy et al., 2018a shows findings about model configurations hold. Models using acoustic features perform better than those that do not; acoustic features at the 10ms timescale gives better performance than 50ms; incorporating linguistic features gives improvement over using acoustic features at 10ms alone; and doing so at a separate timescale brings a greater improvement. Loss (Figure 5.2) shows a similar pattern, although losses for the reproduced experiment are much closer to the original, and sometimes indicate better performance in the reproduction.

The overall patterns also hold on the other data-sets; looking at the correlation between the same metric across data-sets, most are above 90%, with a p-score <0.01 on both Spearman and Pearson.[1] Results on Switchboard always out-perform other corpora. This may be because the linguistic features are more useful than those from Maptask; although linguistic-only Switchboard models are worse than acoustic-only ones, the difference in performance is less pronounced. Also, linguistic-only models trained on Switchboard outperform the acoustic-only models trained on Maptask. Despite having the largest amount of data, the Combined models always perform in between the corresponding Switchboard and Maptask models.

Full results are in Figures A.1 and A.2.

[1] Exceptions to this: F1 500 and F1 onset, on Switchboard v. Roddy and Switchboard v. combined; F1 500 (Pearson only) and F1 onset on Switchboard v. Maptask.

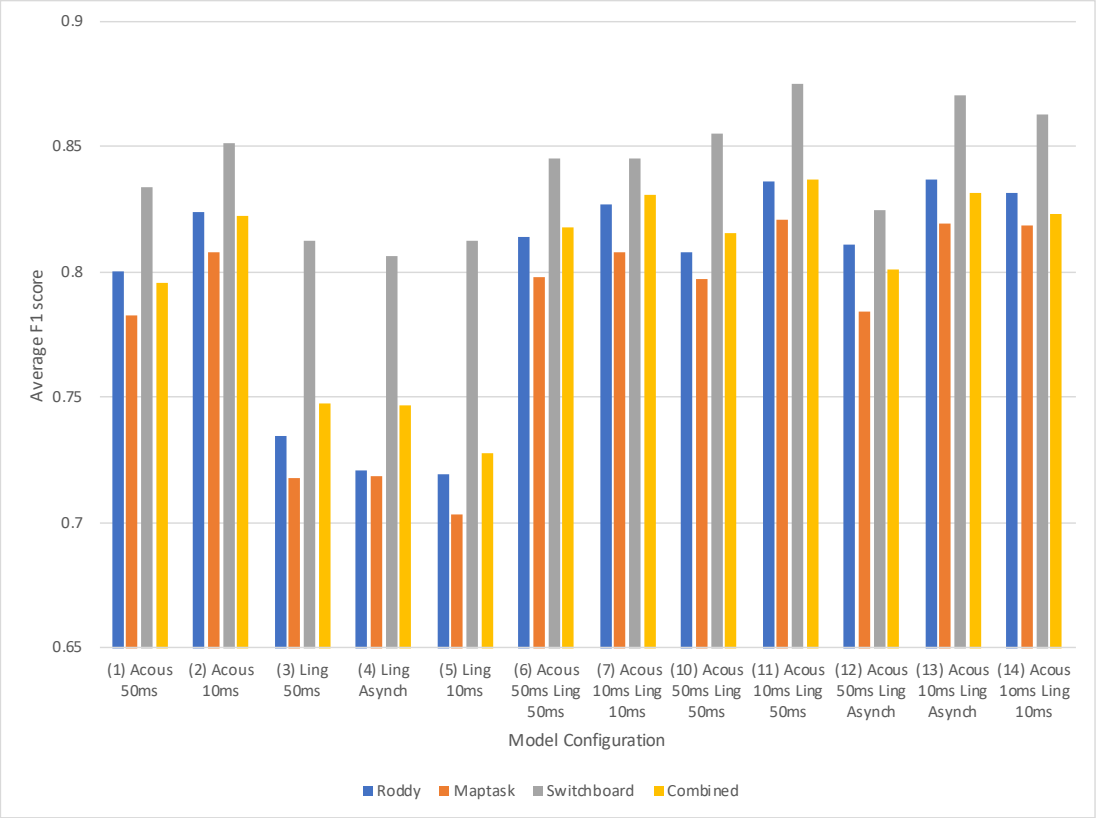

Figure 5.1: Average F1 of various model configurations, trained on different corpora: Maptask results from Roddy et al., 2018b (Roddy), my reproduction (Maptask), and results on Switchboard and Combined.

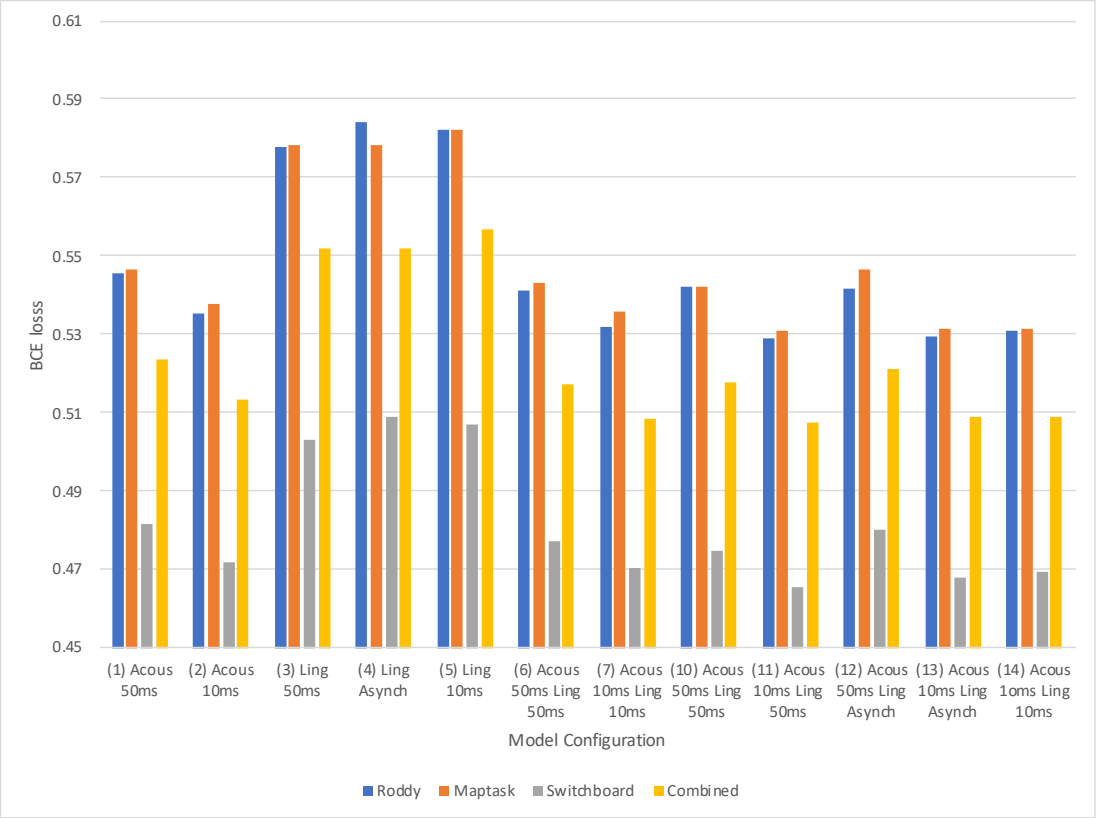

Figure 5.2: BCE Loss of various model configurations, trained on different corpora: Maptask results from Roddy et al., 2018b (Roddy), my reproduction (Maptask), and results on Switchboard and Combined.

5.3 Discussion

This experiment shows that more data is not always better; if it were, Combined models would perform best. However, it is important to consider how these results would look if we ran a different set of experiments. If we only added the Switchboard data to our Maptask set, it may appear to have improved overall results. In reality, it may not be improving performance on the Maptask data, but performing well on Switchboard. This highlights the importance of a consistent test set. I explore this further in Chapter 6.

Again, more hours of Maptask training data do not equate to better evaluation scores; Switchboard models perform better than Maptask models. This is surprising considering the Maptask conversations are highly constrained. I expected the narrow task to reduce variation, compared to the more open-ended Switchboard conversations, especially since Switchboard covers varying topics. In addition, recording conditions in Maptask are higher quality, with Switchboard recorded via telephone.

If Heeman and Lunsford, 2017 are correct, speakers in Maptask are optimising for task completion. Perhaps they make unpredictable interruptions or pauses to think, as necessary to complete the task efficiently. In contrast, Switchboard speakers have less clearly-defined goals,[2] so are perhaps more free to minimise gaps and overlaps. Small talk, as Switchboard could be characterised, does have structure, so although the topics in Switchboard vary greatly, the pattern of turn-taking may be fairly predictable. Gilmartin and Campbell, 2014 believe that chat has a less nested structure than task-based dialogue, so it may be easier for models to identify turn-taking patterns.

[2]Switchboard speakers are given *some* objectives for the conversation, such as "Find out the other caller's favorite pro baseball team and where it's headed this year?" (Godfrey and Holliman, 1993)

6. Experiment: Generalisation Across Corpora

To understand how well models are really generalising – learning true turn-taking cues – rather than just memorising properties of the data they are trained on, I test cross-corpora. If models achieve good performance on test sets from other corpora, they may be learning to generalise. However, I would expect best results when training and testing on the same corpus. Testing models trained on Combined against other test sets can also tell us more about what models are learning. I expect the Combined training set may slightly improve test results for Switchboard, over models trained on Switchboard only, because although Maptask is a constrained task, some of the interactions contained with it are likely to be relevant to predictions that need to be made on Switchboard. However, I hypothesise that it will make results worse on the Maptask test-set; adding data from a more open-ended domain will be less relevant, and may confuse the model.

I expect models using acoustic features to generalise better across corpora, than models using linguistic features only. This is because linguistic features are probably more task-specific. For example, in Maptask some place names on the maps given to participants – such as "Baboons" – might provide turn-taking cues specific to the task and to the context, that are unlikely to appear in Switchboard. If these words do appear in Switchboard, they are probably used very differently, so are unlikely to be predictive.

6.1 Method

I use one model configuration from each modality. This was straightforward for acoustic-only models, but required additional data processing for models using linguistic features.

Roddy et al., 2018b do not deal with unknown words, which does not allow for cross-corpus testing using linguistic features. I train new models for modalities with

	Types frequency 1			Types frequency <= 5		
Dataset	Count	% of total types	% of total tokens	Count	% of total types	% of total tokens
Maptask	560	32.20	0.40	941	54.11	0.67
Switchboard	2766	45.44	2.20	4742	77.90	3.77
Combined	2933	42.97	1.10	4988	73.07	1.87

Figure 6.1: Count of low frequency types by corpus. Figures suggest less variation in words used in Maptask, compared to Switchboard, as expected from the constrained nature of the Maptask task. Statistics prepared using code adapted from Smith, 2020.

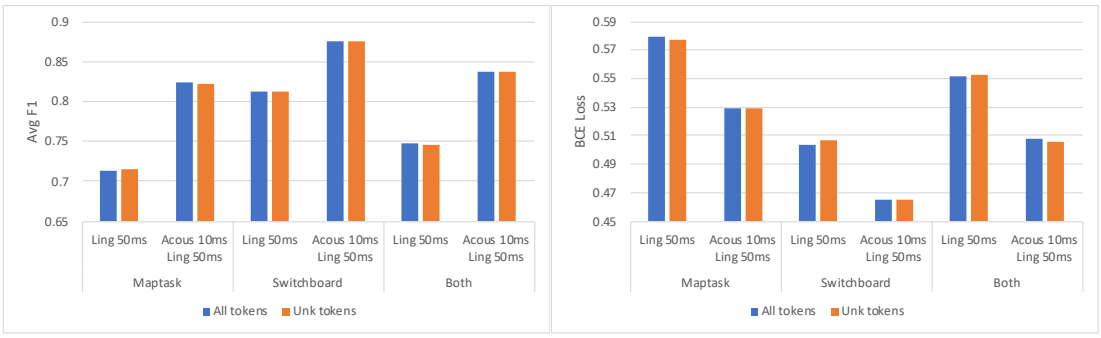

Figure 6.2: Average F1 (left) and BCE loss (right) of models trained using all tokens, and with unknown word tokens replacing low frequency tokens.

linguistic features, replacing tokens of frequency 5 or less with unknown word tokens. Compared to a threshold of 1, this greatly reduces the vocabulary the model must learn, whilst still removing a small percentage of total tokens (see Figure 6.1). This makes little difference to model performance (see Figure 6.2; full results in A.3 and A.4). This is probably because the model mostly learns from acoustic features, and because low frequency words were not providing much predictive information.

Against each model configuration, I ran the test sets from the other corpora.

6.2 Results

Results on the original models reported here, and in future chapters, vary slightly from those in Chapter 5; I am now using only models I trained myself, my colleagues' models being unavailable. Figure 6.3 shows that for models trained on Maptask or Switchboard, performance is generally best on the corresponding test set. An exception is the linguistic-only configuration trained on Maptask, which has a higher average F1 on the Switchboard tests, although the loss fits the general pattern. This may be because some embeddings in Maptask models are over-fitted to the training data. Whilst these words

may also appear in the Maptask test set, they are likely to appear less, or not at all, in Switchboard, so the effect of the over-fitted embeddings is not seen as dramatically in the Switchboard test results.

For models trained on Combined, however, loss and average F1 on the Combined test set falls between the loss on the other test sets for all model configurations. Rather than generalising, the model seems to perform very differently on Maptask and Switchboard; the Combined test results conceal that. Maptask data seems to be more challenging than Switchboard data; loss is always lower and F1 scores are always higher for Switchboard-only models than Maptask-only models. We can also see in Figure 6.3 that training on Combined has made results on Maptask slightly worse for acoustic-only, and markedly worse for the other configurations. Loss follows a similar pattern. On the other hand, training on Combined makes little difference to Switchboard test results, except a small decline in performance for the bi-modal model. Testing for correlations within training sets, but across model configurations, there is no significant correlation (Pearson or Spearman) between model results for each training set, using any metric, apart from between models trained on Switchboard and on Combined, where there is a Spearman correlation of 87% or above for all metrics ($p<0.01$). This suggests Combined models primarily learn from Switchboard data.

The hypothesis that performance will be best when training and test sets match is largely true, although the Combined corpus acts rather differently. The hypothesis that training on Combined would improve results for Switchboard, and worsen results for Maptask does not hold; it made little difference to Switchboard, but did worsen most Maptask results.

Finally, regarding the hypothesis that acoustic features would generalise better than linguistic features, Figure 6.4 is inconclusive. For models trained on Switchboard, metrics do show the smallest cross-corpus decline for acoustic-only models. This does not hold for models trained on Maptask where, as previously discussed, Average F1 actually improves and BCE Loss only declines slightly for linguistic-only models, when tested on Switchboard. For the other model configurations, all metrics get markedly worse.

Full results are in Figures A.5 and A.6.

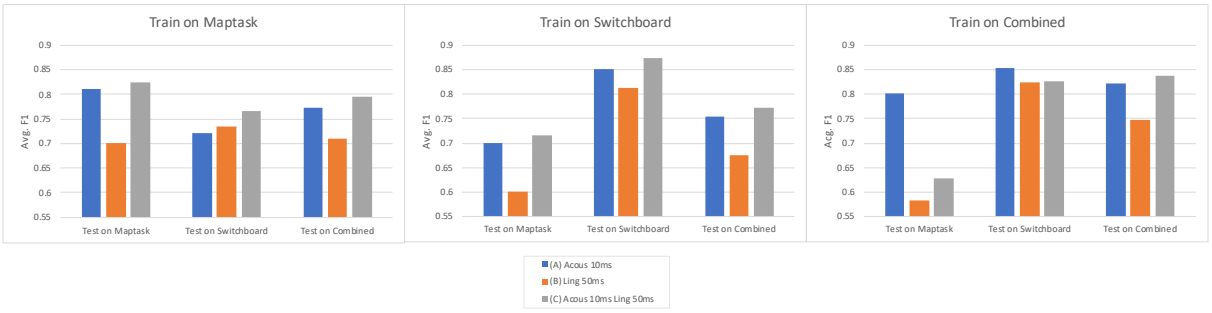

Figure 6.3: Average F1 for cross-corpus tests i.e. trained on one corpus and tested on another.

	Change to BCE Loss		Change to Avg. F1	
	Train on Maptask	**Train on Switchboard**	**Train on Maptask**	**Train on Switchboard**
(A) Acous 10ms	+15.32%	+28.03%	-10.95%	-17.73%
(B) Ling 50ms	+2.49%	+37.31%	+4.60%	-26.09%
(C) Acous 10ms Ling 50ms	+10.41%	+28.97%	-7.08%	-18.12%

Figure 6.4: For Maptask and Switchboard, percentage change of evaluation metrics from original test set, to test set from the other Corpus: for Train on Maptask, change in metrics is from Test on Maptask to Test on Switchboard; for Train on Switchboard, change in metrics is from Test on Switchboard to Test on Maptask.

6.3 Discussion

The results on models trained on Combined suggest that, rather than treating the data as a single body of evidence, the model may be identifying differences between corpora (e.g. from recording conditions) then learning a separate model for each: not generalising across corpora. Whilst this is not surprising, it is somewhat disappointing; it demonstrates that this architecture has a limited ability to learn from the data that is available. We would like our models to mimic humans, who can learn fundamental cues and rules of turn-taking across different dialects and audio conditions (e.g. over the phone, face-to-face). Changing model architecture could help. For example, a domain-adversarial model could encourage models to ignore fundamental differences between corpora, and generalise better, in analogy with Liang et al., 2019's model for cross-culture emotion recognition. Alternatively, inspiration could be taken from Huang et al., 2013's multilingual acoustic model using data from multiple corpora – or languages – simultaneously to train the internal layers of a model as a feature extractor, but maintain a separate output layer for each corpus.

Acoustic-only models perform almost as well as models trained on both acoustic and linguistic features. From an implementation perspective, this experiment showed me it is not straightforward to use linguistic features across corpora. Additionally, if we rely on manual annotations, annotation differences between corpora may cause problems, whereas with automatically extracted features we can at least be sure they are consistent. For live dialogue systems, turn-taking modules would need to rely on ASR outputs, which will also perform differently on different data. In both cases, acoustic features enable good performance, for a less complex data pipeline. This experiment showed that acoustic features may also generalise better. This is true on Switchboard, and may hold for other conversational speech. The results on Maptask may be the result of over-fitting on certain recurrent task-based words, that simply do not appear in Switchboard. This is likely due to the distinction between conversational and task-based speech; different modalities may behave differently for different domains.

One drawback of this experiment is that it is not well controlled. There are many differences between the corpora, for example: dialect, recording quality, and task. From a purely scientific perspective, it would be useful to isolate these variables to understand how they individually affect the ability of models to generalise. However, from an engineering perspective the data available to us when we build real-world systems often has similar drawbacks; it is not feasible to collect the perfect data from scratch for every system, so we should explore ways to make the best of the data we have. Since this is likely how many systems are being trained anyway – flippantly, by chucking in as much data as we can find, and hoping for the best – it is pertinent to explore what happens when we do so.

7. Experiment: Effect of Roles

In Chapter 6, the model appears to learn different rules for different data sets, rather than generalising; I wondered if there may be similar intra-corpus effects. I explore this in terms of speaker roles, but it could also be examined in terms of socio-linguistic factors.

Maptask has strictly defined roles, which are likely to exhibit differing turn-taking behaviour, so I would expect a difference in model performance on the *Instruction Giver* (G) and the *Instruction Follower* (F). From listening, G usually talks more, so performance is likely higher on G than F. For Switchboard, I do not expect a role difference, since A and B are given the same instructions and objectives for the conversation.

If a role difference is found, a model trained specifically on each role may improve performance for speakers in that role. However, it may have the opposite effect, because doing so reduces the quantity of training data.

7.1 Method

I examine existing performance differences by role for models trained in Chapter 5, looking at Mean Absolute Error per speaker.

I also train models on individual roles. I test models on data from the role they were trained on, the other role, and the original, complete test set.

7.2 Results

Hold-shift predictions are not possible when testing on a single role, since these rely on comparing probabilities for each speaker. I report only loss and F1 Onset in these cases.

Figures 7.1 shows a clear difference in performance between Maptask roles, for all configurations; Mean Absolute Error (MAE) is lower for G than for F, in line with

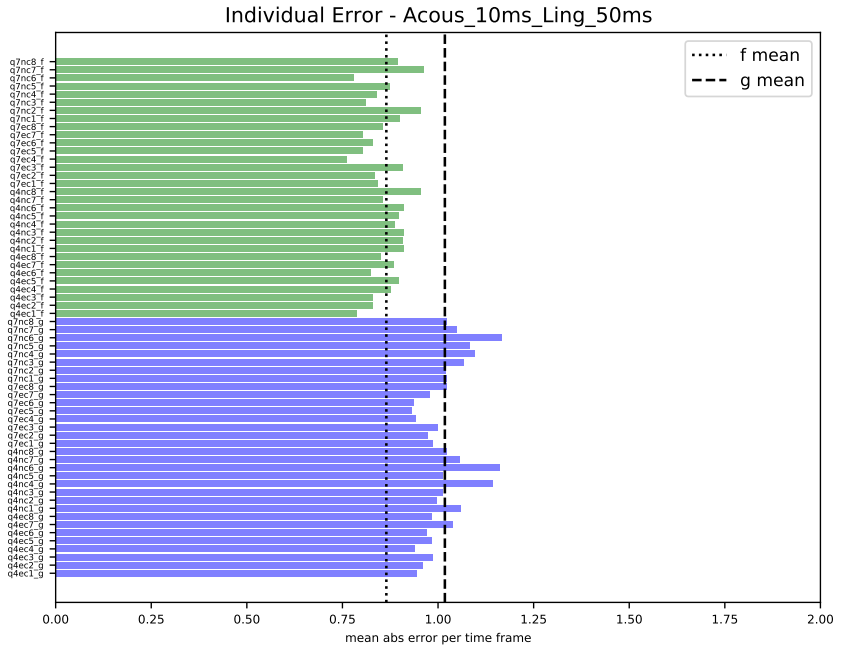

Figure 7.1: MAE per person for bi-modal models on Maptask. F is in green, G is in blue.

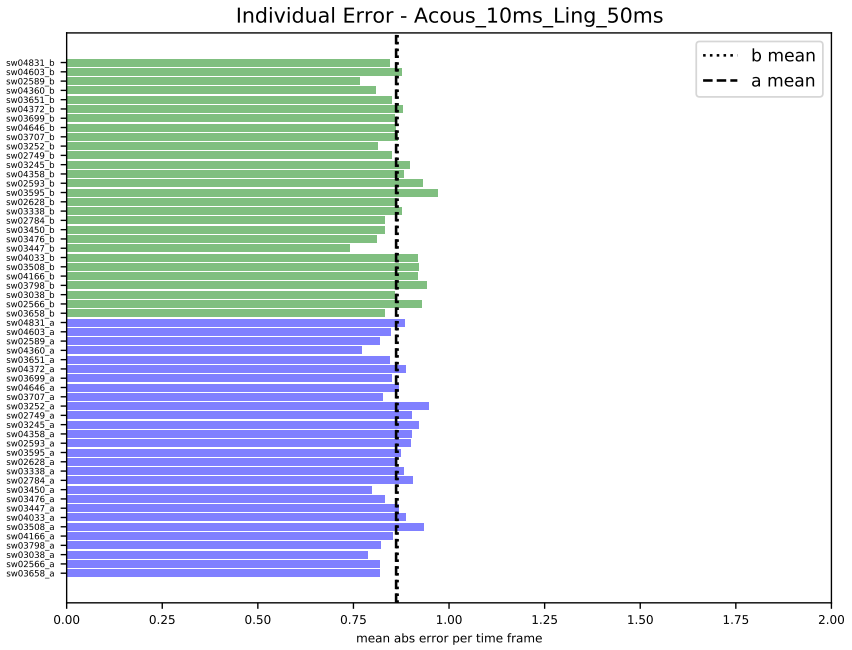

Figure 7.2: MAE per person for bi-modal models on Maptask. B is in green, A is in blue.

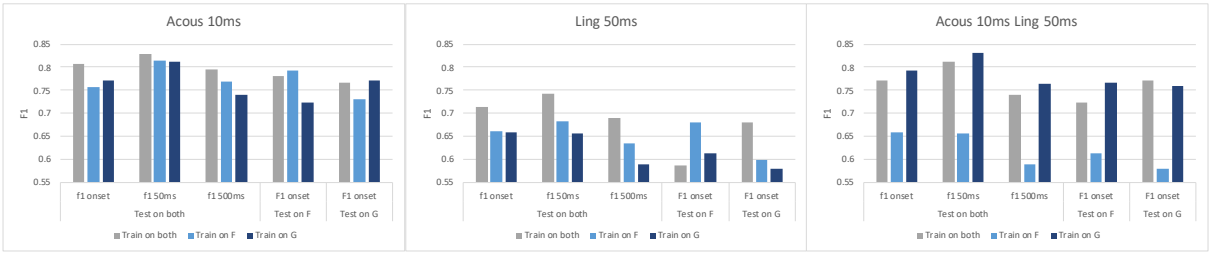

Figure 7.3: F1 scores of role experiments on Maptask.

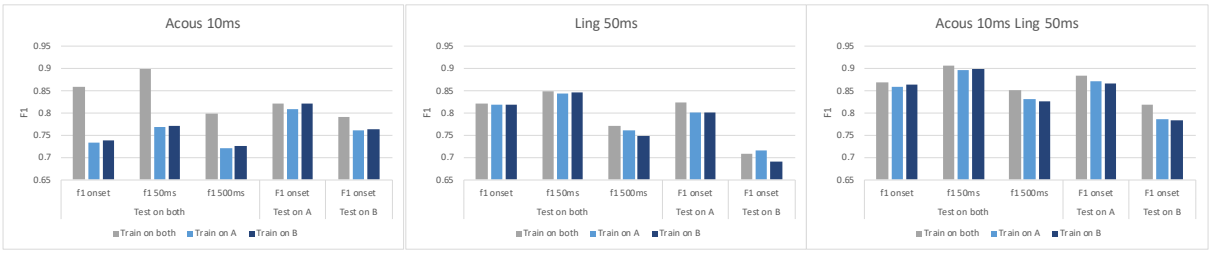

Figure 7.4: F1 scores of role experiments on Switchboard.

my hypothesis. Figure 7.2 shows, also as expected, a negligible difference between Switchboard roles. Other modalities were consistent with this (see Figures A.9 to A.12.)

Results for models trained on different Maptask roles (Figure 7.3) show that for acoustic- and linguistic-only models, performance is almost always better when trained on both roles. This suggests the model is not confused by role differences in the data, and can generalise some of what it learns across roles. Testing on F, however, gives better F-scores when trained on F-only. The bi-modal results are more surprising; best F1 scores for all test sets are from models trained on G only, except the G-only test set, which performs slightly better on the model trained on both roles. Bi-modal models trained on F-only perform well below the other models, and are similar to or worse than the F-only linguistic-only model configuration.

Results for models trained on Switchboard roles (Figure 7.4) show less variation, unsurprisingly given no role difference was found. Rarely, models trained on one role slightly outperform those trained on both. The biggest effect of using less data is in acoustic-only models; perhaps models continue to learn from additional acoustic data, but may be approaching a performance ceiling using word features only. Bi-modal modals do not show such performance differences, suggesting that when less acoustic data is available to the model, it can compensate using the linguistic features.

Full results are in Figures A.7 and A.8.

7.3 Discussion

Whilst the role distinction in Maptask may seem artificial, in our daily lives where we hold different roles in different interactions, and this affects our behaviour. These roles likely interact in complex ways with facets of our identity such as gender, age and race. Examining role differences is a starting point to exploring this, but there is considerable scope for further work.

Stivers et al., 2009 find dialogue act predictive of turn-taking behaviour, and Aldeneh et al., 2018 find model performance varies by dialogue act. Both roles in Switchboard probably use a similar distribution of dialogue acts, since all speakers are engaging in conversational speech. Maptask role differences could be explained by a different distribution of dialogue acts used by each role; from listening, F mainly utters back-channels and short responses, and G utters more informative statements. It would be interesting to examine further the relationship between roles, dialogue acts and model performance.

As well as variation per role, there is marked variation for individual speakers, even within roles, especially in Maptask. I listened to a selection of speakers from each Maptask role, with notably good or poor model performance. For G, models perform well on slow speakers who are strongly focused on the task (see Figure 7.5). For F, the models perform best on speakers who mainly utter short answers and back-channels (see Figure 7.6). In these dialogues, G is clearly driving the conversation; F occasionally asks for clarification or proactively volunteers information, but even then are concise. In contrast, for both F and G, speakers who experienced lower model performance were more chatty, built up rapport and seemed at ease (see Figures 7.7 and 7.8). These participants speak quickly with a broader range of dialogue acts. They stop to laugh at mistakes, and make comments not directly related to task completion. This chimes with Cassell et al., 2007, who tell us that when speakers build up rapport they treat turn-taking less rigidly, and are more likely to cover more topics.

```
G: starting above the site of the forest fire
F: certainly am
G: right
G: go east west sorry go west slightly just past it and
    down and then you go south
F: and go hang a big left right past it
F: right
G: so you go so you're going east west then south down
F: yep
G: have you got a picnic site
F: nope
```

Figure 7.5: Example dialogue where there is high performance on the G speaker (from Maptask q7ec5). Utterances are focused on the task: generally simple questions or imperatives.

```
G: well say the top's north and
F: right
G: go west
G: eh
F: okey-dokey
G: until you've just got past the edge of the crest falls
F: right
```

Figure 7.6: Dialogue extract where the model performs well on F (from Maptask q7ec4). Utterances are short, simple and there is little rapport.

```
G: yeah you go right around the fields
G: and then you go left again to the base of your thumb
F: right okay right i'm just doing that now
G: 'cause it looks like it looks like a hand sort of well
    my thumb's a lot bigger than yours
F: god i didn't realise my thumb was such a funny shape
F: yeah right it's okay we've handled it i think
G: okay so
G: now
F: it's mental shape that
G: i know it's cool
```

Figure 7.7: Dialogue extract where the model performs poorly on F (from Maptask q7nc2). F makes long, off topic utterances and displays rapport with G.

```
G: have you to the left of that down a bit have you got baboons
F: no
F: i don't have anything
G: right
G: well we what we have to do is like draw like a cup
    shape thing round if you go down about an inch and
    along about no down about an inch and a half along an
    inch and a half and then up
F: up and then a cup
G: an-- in a cul-- in a kind of you know bowl-shaped thing right
F: okay
```

Figure 7.8: Dialogue extract the model performs poorly on G (from Maptask q7nc6). G speaks quickly and gives long, creative explanations.

One female speaker with a working-class Scottish accent appears repeatedly in the conversations I examined for their high error scores: F in Figure 7.7 and G in Figure 7.8. She sounds nervous but is enjoying the task, swearing and making jokes to build rapport with the other speakers. In contrast, speakers I examined for good model performance are often male, or middle-class female speakers. This suggests that speakers from less privileged groups are less well served by these models. Lower performance speakers in Switchboard have strong Southern accents, and seem to demonstrate higher rapport, which also supports this suggestion.

8. Experiment: Exploring Prediction at Onset

Whilst Prediction at Pauses (Section 4.4.2) has been commonly modelled, Skantze, 2017 introduces the Prediction at Onset metric (Section 4.4.3).

Since Prediction at Onset measures something different to Prediction at Pauses, I expect models to have differing success for each measure. However, for Maptask models, correlations (Pearson and Spearman) between Prediction at Onset and other evaluation measures is always greater than 95%, with $p<0.01$. Correlations for Switchboard (Figure 8.1) are not as high, but show moderate correlation. The Combined corpus has higher correlations than Switchboard, likely the influence of the Maptask data.

Given the difference in correlations between corpora, I wanted to understand if metric parameters are domain specific. If so, they are potential levers we could use to adapt the model without retraining neural network parameters. They are also relatively interpretable. However, considering the way *long* and *short* are defined, I do not expect the predictive information contained in each second of output predictions, or the classification thresholds to vary much by modality or corpus, because the definition of each class does not change.

For *long* utterances I expect the model to predict a high probability of voice activity for at least the first two seconds after the prediction point (shown in Figure 8.2), followed by either voice activity or silence in the third second. For *short* utterances, I expect the model to predict an initial 1s or less of voice activity, followed entirely by silence for the remaining frames. Therefore, I do not expect the first second of voice activity to be useful, since in both classes of utterance voice activity should be predicted. I expect the 2nd second to provide more predictive information: the speaker should be predicted to stop speaking during the 2nd second in a *short* utterance, but must continue speaking in a *long* utterance. I expect the third second to be similarly predictive; in a *short* utterance the speaker must stop speaking by this point, and in a *long* utterance they should continue speaking for at least 500ms.

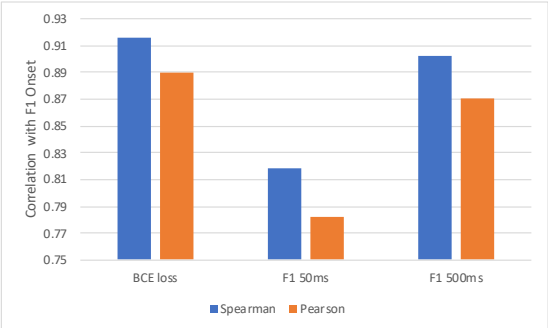

Figure 8.1: For Switchboard results from Chapter 5, correlation between Prediction at Onset and other measures. All p-values $<$0.01.

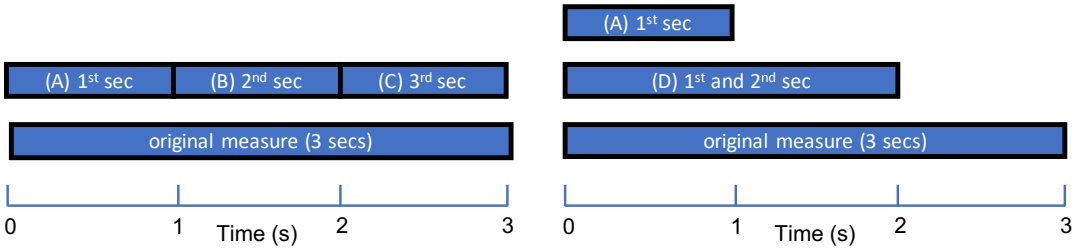

Figure 8.2: The original onset prediction metric uses the average of all 3s (60 frames) of voice activity prediction, (see Section 8). I examine how each second contributes individually to prediction accuracy (left), and (right) whether additional seconds provide useful information over the previous second(s) of predictions.

8.1 Method

Figure 8.2 shows the modified Prediction at Onset metrics I tested on model configurations trained in Chapter 5. I did not retrain models or change their architecture: they still output 60s of predictions.

8.2 Results

Figure 8.3 shows using all 3s of output predictions generally gives the best performance for Prediction at Onset; there is useful information throughout the predictions. My hypothesis that seconds 2 and 3 would be more predictive is incorrect. The exception is Switchboard Acoustic 10ms, where incorporating predictions after the first second harms F1. This contradicts my hypothesis that results would not vary by corpus. The pattern of results suggests speech activity is easier to predict over the shorter term, with

the 1st second of predictions giving better results, likely due to more accurate voice activity predictions. There also appears to be a difference between modalities, which I did not expect. For acoustic-only models, the first second of predictions is clearly the most useful. For linguistic-only models this holds, but barely, on Switchboard. Other corpora show different patterns. For Maptask, predictions improve in later seconds; on Switchboard there is little difference in performance between each individual second; and on Combined the 2nd second gives the best performance after the original metric. This suggests that linguistic features will not allow models to generalise across corpora as successfully as acoustic features; words appear to be providing very different information in the different data-sets.

Figure 8.4 shows that with each additional cumulative second of output predictions used, F1 scores almost always improve, again with the exception of Switchboard Acoustic 10ms, which performs slightly better without the third second of predictions. There is, again, a difference in the linguistic-only models; on Maptask predictions improve markedly with each additional second of model outputs, whereas Switchboard scores are relatively stable, reinforcing the idea that the linguistic features are providing different cues in each corpus.

Classification thresholds (Figure 8.5) also vary by corpus, model configuration, and metric version, counter to my hypothesis. In particular, the difference between linguistic-only models for each corpus are evident. For most model configurations, output predictions must show a considerably higher probability of voice activity in the first second than in other seconds to be classified as a long utterance.[1] The threshold lowers in second two, then lowers further in second three. Again, the Switchboard linguistic-only model is an exception, with the threshold remaining very low throughout. This suggests that, considering linguistic features alone, the model finds *short* utterances extremely unlikely, even when the probability of voice activity is low throughout the utterance. This is surprising considering there is no such pattern for other modalities, so it seems unlikely that the probability of *short* utterances is truly low in this corpus. It is also unexpected that the bi-modal model trained on Combined has such a high threshold in second three, considering that this is not the case on the other two data-sets, from which Combined is constructed.

[1] Recall that utterances falling **above** the classification threshold are predicted to be *long*; all others are predicted to be *short*.

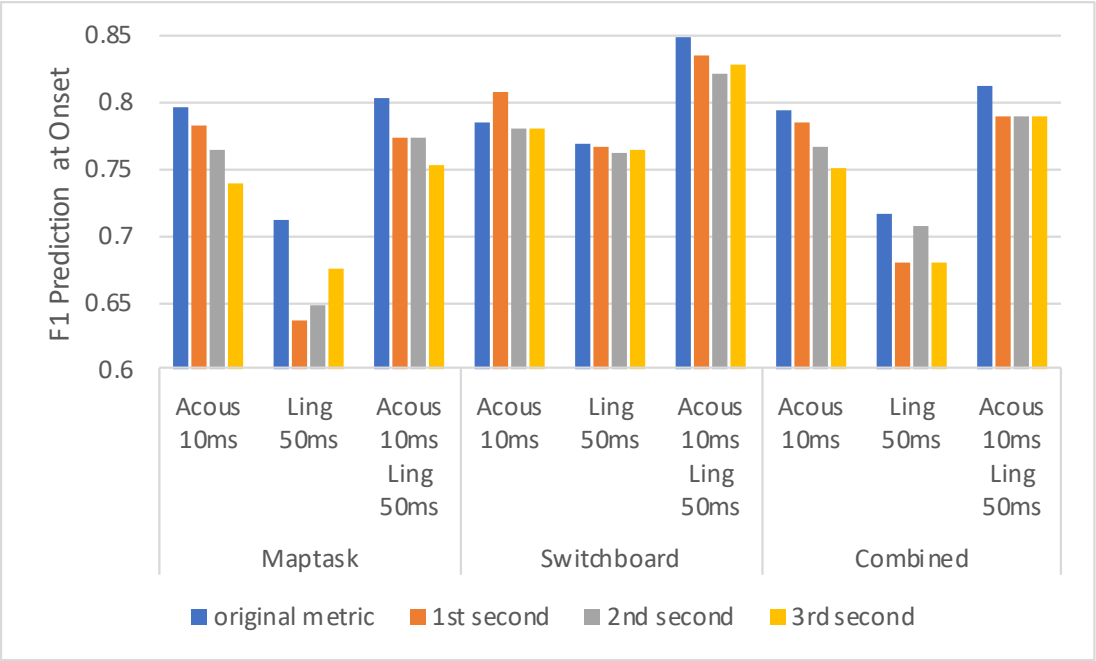

Figure 8.3: F-scores for Prediction at Onset, using the original measure, and individual seconds of model outputs.

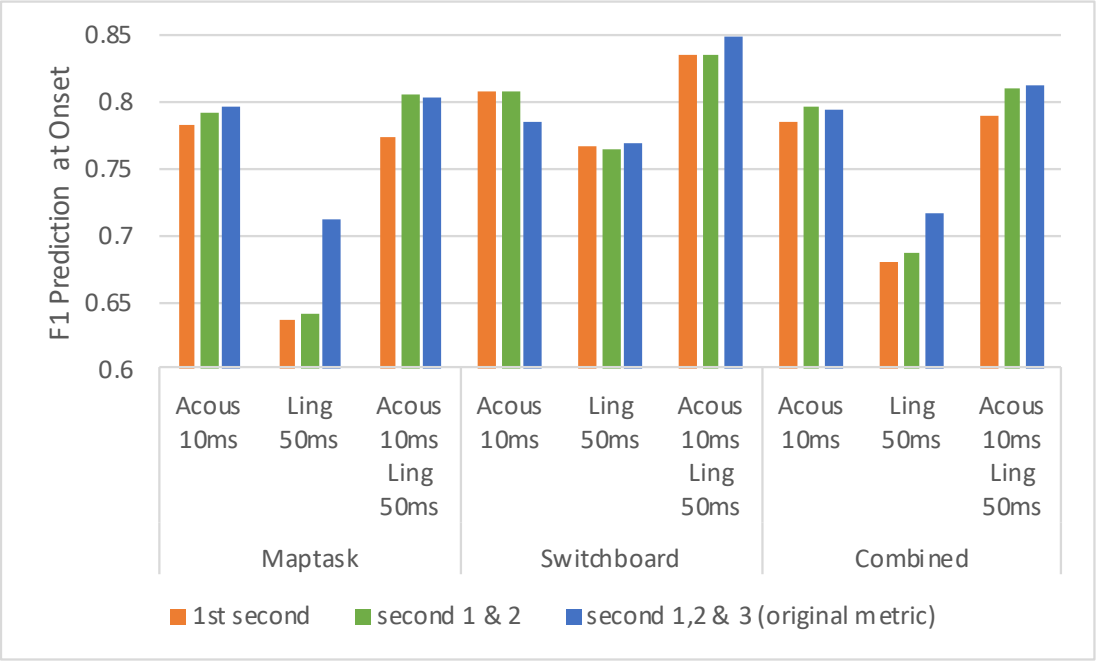

Figure 8.4: F-scores for Prediction at Onset, using increasing number of seconds of model output.

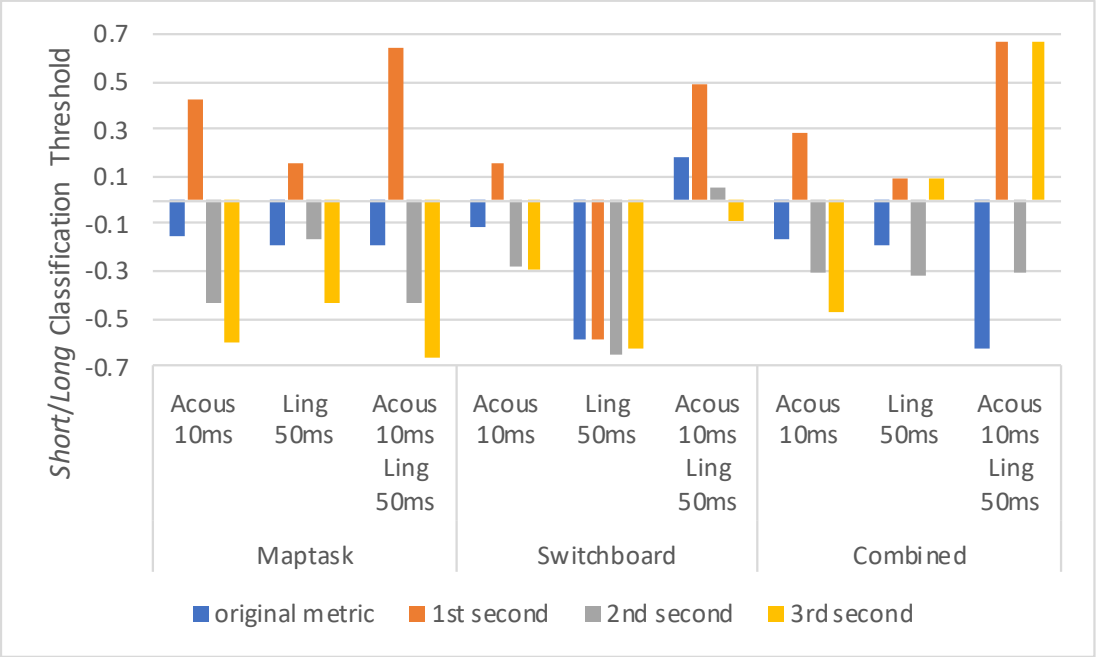

Figure 8.5: Classification thresholds for Prediction at Onset, using the original measure, and individual seconds of model outputs.

8.3 Discussion

It is surprising that using just the first second of voice activity prediction gives most of the information needed to classify *short* and *long* utterances. Results suggest all of my hypotheses are incorrect. Model outputs are probabilities of voice activity for each frame, but when they are used to make utterance length predictions, they do not appear to be used as such. Perhaps this is not a problem if the classification can produce good[2] results; figures A.1 and A.2 show that the lowest F1 Onset is 0.7167 for the model configurations we are considering here, and reaches 0.9046 with the bi-modal model on Switchboard. However, it is concerning that the metric is not behaving as expected; do we really know what we are measuring? Understanding our metrics should be a priority if we want transparent results, particularly when using black-box models.

Results show that, in some cases, we may be able to improve model performance on this task simply by changing the input we use for the classification, for example in the case of the acoustic-only Switchboard model. If our goal was to maximise performance, does this mean we should tune the metric on every data-set? It may also be helpful to tune the definitions of *short* and *long*. This could be viewed as "cheat-

[2]Though what is *good* is challenging to define.

ing", and make it challenging to compare results. However, it is also likely that back-channels do vary in duration, for example, in different tasks or dialects, so altering the definitions could be well-motivated, perhaps through analysis of data with dialogue act annotations. In terms of generalisability, tuning may or may not be desirable. On one hand, my intuition is that extensive tuning will decrease the application of the model outside the context for which it is tuned. On the other hand, it could be a quick way of adapting models to new contexts; models could be re-tuned without needing to be re-trained, which would be especially advantageous if computationally-expensive pre-trained models became available for this task. In this way, we would not be creating a single general model, but in some sense the model would be generalising, because we could use it with greater success in more contexts.

We have seen that hold-shift and onset predictions are often correlated, but they may not be for all corpora or tasks. If we find that performance between hold-shift and onset is very different in certain circumstances, what should we optimise for? Such a decision depends on the intended use of the model, and may impact different speakers in different ways; it is likely that speakers taking on different roles, or from different groups utter back-channels with differing frequencies. It would be difficult to judge the impact of this on end-users without extrinsic evaluation. If this is done, care should be taken to select a diverse user group. Certainly, what and who we decide to measure will make a difference to what we find.

9. General Discussion

I revisit the aspects of generalisability outlined in Section 1.1, in view of my results.

9.1 Mechanical Generalisability

Mechanical issues can generally be overcome, though not necessarily straightforwardly. They should be considered early on when developing turn-taking models. For example, to perform cross-corpus tests in Chapter 6, I had to train new models, with an unknown word token. This made little difference to performance, possibly because there was no real pattern as to which type of words were unknown, but may have affected results for individual speakers; mechanical decisions may impact other aspects of generalisability.

Future work could explore training separate unknown tokens for different parts of speech, which may improve model performance on unknown words, although this would complicate the data pipeline with potentially little difference to overall F-score. However, I would expect under-represented groups in the data to use more unknown words than highly represented groups. Therefore, how we deal with low frequency or unknown words is likely to have an impact on how well models serve these groups, and how well they generalise.

9.2 Intra-Corpus Generalisability

My experiments show that models learn turn-taking cues more successfully on some data than others: conversational speech appears to be easier than task-based speech, and models perform best when speech is slow and deliberate, with less rapport between speakers. Some dialogue acts are also more predictable than others. I found considerable variation between speakers in all corpora, and between defined roles.

Turn-taking behaviour seems hardest for models to predict in speakers that are

having most fun, and being creative; by definition, creativity is unpredictable. My experience is that I must be serious and focused to use current dialogue systems. Will we ever be able to have a light-hearted, fun conversation with a dialogue system?[1] The light-hearted conversations in Maptask were all conversations where speakers could not see one another, whilst most of the conversations where the model performed well were conducted with eye-contact. Perhaps lack of eye-contact made speakers feel more at ease, or perhaps they were communicating more emotion via speech, in the absence of visual signals. Since performance is better on Switchboard, where all speakers are strangers, without eye-contact, it is likely that it is familiarity that is making speakers perform less predictable turn-taking behaviour. Certainly, from listening to Switchboard conversations, they are generally quite formal. Like in Maptask, speakers who are more animated and less formal experience worse model performance.

9.3 Cross-Corpus Generalisability

Models are not necessarily able to generalise across data from multiple corpora; they may be learning separate models for each. Using overall F1 scores does not give us this picture. we need more nuanced evaluation measures: we could report metrics on different roles or groups within the data. Neural networks are useful because they identify patterns and groups in data, but may also use patterns we have not considered, in potentially problematic or discriminatory ways. Although it would be impossible to comprehensively represent all axes of variation, attempting to capture this in some way would give us a more rounded view of model performance.

Cross-corpus generalisability also applies to corpora in different languages. Brusco et al., 2017 suggest a model trained on another language could be a starting point for a turn-taking module on a new language. Although, mechanically, we could use ASR and machine translation to enable linguistic features to be used cross-lingually, my results suggest that acoustic features generalise better, and will likely be more applicable in a cross-lingual context. I also expect acoustic features to be more useful to train multilingual models.

Roddy et al., 2018b also perform experiments using visual features (gaze direction) which, combined with acoustic features, gives a small, significant improvement over acoustic features alone. Notwithstanding considerable mechanical complications, I believe gaze features could generalise cross-corpus in a similar way to acoustic fea-

[1] If that is what we want.

tures,[2] since they are also a closed set of continuous features. In contrast, words are represented as a closed set of embeddings (vectors which must be trained jointly with the network). Since different corpora have different sets of words, and use words differently, these do not generalise well. Acoustic and visual cues are also both controlled less consciously than word choices, which may be related to how they generalise.

9.4 Adaptation v. Generalisation

As Futoma et al. discuss, we should not write off models that do not generalise; they may still be useful in some circumstances. In addition, my experiments have shown there may be simple ways we can adapt models to perform better in new contexts. For example, as explored in 8, changes to which output predictions are used to make the decision sometimes improves model performance. The prediction threshold could also be tuned manually, or by using a R.O.C curve on different data. Neither of these would require any additional training of network parameters, so are computationally cheap. This could be important if heavy-duty pre-trained embeddings became available (e.g. in the style of BERT). However, it is prudent to understand the workings of any metrics further before attempting to manipulate them in this way, so we understand the implications of any changes.

There appear, then, to be two routes to pursue: developing models which generalise better – for example by investigating new model architectures (see Chapter 6) –, or developing adaptation techniques that allow models to be used more successfully in specific contexts. I believe both are important, and future systems are likely to use a combination of approaches, especially since it is not clear that a perfect general model for turn-taking is possible. In addition, both techniques taken to the extreme seem to lead to similar results: the most general model is in some sense the one that can best adapt to individual speakers.

[2] although perhaps not in conversations where there is no eye contact between speakers

Until we reach this probably unachievable goal, we should consider the advantages of pursuing generalisation or adaptation for a given system. Could we train or adapt various models for different contexts, for example if we already have access to information about a user's role? Or do we know little about our users, in which case aiming for as general a model as possible may be the best option. If we set too broad a scope for generalisations, will our models give a mediocre performance on everyone? This may be a *fair* outcome, but probably not a desirable one.

10. Conclusions

I have explored various aspects of generalisability by training models on various data sets, and examining the results. I have found:

- Generalisability is a complex concept, and it is just as complex to understand whether our models are doing it.
- LSTM turn-taking models can struggle to generalise across both intra- and inter-corpora effects. Other architectures may improve generalisation.
- Acoustic features provide the most useful generalisable information, although linguistic features may help compensate if there is less acoustic information.
- There is more variation in how useful linguistic features are between corpora and roles, than there is for acoustic features, perhaps because acoustic features are less likely to overfit.
- Speakers who are less formal and develop more rapport may be less well served by turn-taking models. For the corpora I used, this may also be true of speakers with less prestigious accents.
- We need to better understand the meaning and limitations of our evaluation metrics, and develop more nuanced ways of evaluating model performance – both intrinsic and extrinsic – beyond chasing increases in F1.
- There may be parameters of our metrics we can tune for "quick wins" in adapting models to new contexts, if they are not able to generalise.
- Extrinsic evaluation is necessary to understand what effect changes to our model really have on users of dialogue systems

Finally, we should remember that every speaker in our data is a person, not just a number. How well turn-taking predictions work for them will make a difference to how often they are ignored or interrupted by the dialogue systems that will become increasingly ubiquitous as a way to access services in the future.

Bibliography

Aggarwal, C. C. (2018). Teaching Deep Learners to Generalize, In *Neural networks and deep learning*. Springer, Cham. https://doi.org/10.1007/978-3-319-94463-0{_}4

Aldeneh, Z., Dimitriadis, D., & Provost, E. M. (2018). Improving End-of-Turn Detection in Spoken Dialogues by Detecting Speaker Intentions as a Secondary Task, In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings*. https://doi.org/10.1109/ICASSP.2018.8461997

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4), 351–366. https://doi.org/10.1177/002383099103400404

Arsikere, H., Shriberg, E., & Ozertem, U. (2015). Enhanced End-of-Turn Detection for Speech to a Personal Assistant, In *Aaai spring symposium - technical report*.

Bauman, R., & Sherzer, J. (Eds.). (1989). *Explorations in the Ethnography of Speaking* ((2nd ed.,). Cambridge, Cambridge University Press. https://doi.org/10.1017/cbo9780511611810

Brusco, P., Perez, J. M., & Gravano, A. (2017). Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish, In *Proceedings of the annual conference of the international speech communication association, interspeech*. https://doi.org/10.21437/Interspeech.2017-124

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-010-9120-1

Cassell, J., Gill, A. J., & Tepper, P. A. (2007). Coordination in conversation and rapport. https://doi.org/10.3115/1610065.1610071

de Kok, I., & Heylen, D. K. (2009). Multimodal End-of-Turn Prediction in Multi-Party Meetings. *Proceedings of the International Conference on Multimodal Interfaces, ICMI-MLMI 2009*, 91–98.

Dethlefs, N., Hastie, H., Cuayáhuitl, H., Yu, Y., Rieser, V., & Lemon, O. (2016). Information density and overlap in spoken dialogue. *Computer Speech and Language*, *37*, 82–97. https://doi.org/10.1016/j.csl.2015.11.001

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292. https://doi.org/10.1037/h0033031

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE - The Munich versatile and fast open-source audio feature extractor, In *Mm'10 - proceedings of the acm multimedia 2010 international conference*. https://doi.org/10.1145/1873951.1874246

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet. Digital health*, *2*(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

Gilmartin, E., & Campbell, N. (2014). More Than Just Words: Building a Chatty Robot, In *Natural interaction with robots, knowbots and smartphones*. https://doi.org/10.1007/978-1-4614-8280-2{_}16

Godfrey, J., & Holliman, E. (1993). Switchboard-1 Release 2 Manual (LDC97S62). https://catalog.ldc.upenn.edu/docs/LDC97S62/

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development, In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings*. https://doi.org/10.1109/ICASSP.1992.225858

Gravano, A., Brusco, P., & Beňuš, (2016). Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish, In *Proceedings of the annual conference of the international speech communication association, interspeech*. https://doi.org/10.21437/Interspeech.2016-585

Gruzin, E. (2020). *Transition Relevance Place Detection Using LSTM Models of Voice Activity (MSc Thesis - University of Edinburgh)*.

Heeman, P. A., & Lunsford, R. (2017). Turn-Taking offsets and dialogue context, In *Proceedings of the annual conference of the international speech communication association, interspeech*. https://doi.org/10.21437/Interspeech.2017-1495

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings*. https://doi.org/10.1109/ICASSP.2013.6639081

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing (3rd ed. draft, October 2019)*. https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf[accessed2ndOctober,2020]

Liang, J., Chen, S., Zhao, J., Jin, Q., Liu, H., & Lu, L. (2019). Cross-culture Multimodal Emotion Recognition with Adversarial Learning, In *Icassp, ieee international conference on acoustics, speech and signal processing - proceedings*. https://doi.org/10.1109/ICASSP.2019.8683725

Olah, C. (2015). Understanding LSTM Networks [Blog]. *Web Page*. https://doi.org/10.1007/s13398-014-0173-7.2

Roddy, M. (2018). LSTM Turn-Taking Code (GitHub). https://github.com/mattroddy/lstm_turn_taking_prediction

Roddy, M., Skantze, G., & Harte, N. (2018a). Investigating speech features for continuous turn-taking prediction using LSTMs, In *Proceedings of the annual conference of the international speech communication association, interspeech*. https://doi.org/10.21437/Interspeech.2018-2124

Roddy, M., Skantze, G., & Harte, N. (2018b). Multimodal continuous turn-taking prediction using multiscale RNNs, In *Icmi 2018 - proceedings of the 2018 international conference on multimodal interaction*. https://doi.org/10.1145/3242969.3242997

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696. https://doi.org/10.2307/412243

Selfridge, E. O., & Heeman, P. A. (2010). Importance-Driven Turn-Bidding for spoken dialogue systems, In *Acl 2010 - 48th annual meeting of the association for computational linguistics, proceedings of the conference*.

Skantze, G. (2016). Real-time coordination in human-robot interaction using face and voice. *AI Magazine*, *37*(4), 19–31. https://doi.org/10.1609/aimag.v37i4.2686

Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks, In *Sigdial 2017 - 18th annual meeting of the special interest group on discourse and dialogue, proceedings of the conference*. https://doi.org/10.18653/v1/w17-5527

Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain, In *Eacl 2009 - 12th conference of the european chapter of the association for computational linguistics, proceedings*. https://doi.org/10.3115/1609067.1609150

Smith, L. (2020). *Personalizing Models of Turn-Taking with Task-Specific Embeddings (MSc Thesis - University of Edinburgh)*.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K. E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Tice, M., & Henetz, T. (2011). Turn-boundary projection : Looking ahead, In *Proceedings of the annual meeting of the cognitive science society*.

Ward, N. G., Aguirre, D., Cervantes, G., & Fuentes, O. (2019). Turn-Taking Predictions across Languages and Genres Using an LSTM Recurrent Neural Network, In *2018 ieee spoken language technology workshop, slt 2018 - proceedings*. https://doi.org/10.1109/SLT.2018.8639673

Yang, F., & Heeman, P. A. (2010). Initiative conflicts in task-oriented dialogue. *Computer Speech Language*, *24*(2), 175–189.

A. Appendix: Full Results

Figures A.1 to A.8 give full results for experiments described in the main text.

	Roddy et al					Reproduction				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
No Subnets (Early Fusion)										
(1) Acous 50ms	0.5456	0.7907	0.8165	0.7926	0.7999	0.5465	0.7678	0.8125	0.7680	0.7828
(2) Acous 10ms	0.5351	0.8154	0.8428	0.8126	0.8236	0.5376	**0.8056**	**0.8282**	0.7897	0.8078
(3) Ling 50ms	0.5779	0.7234	0.7547	0.7249	0.7343	0.5783	0.7116	0.7522	0.6885	0.7174
(4) Ling Asynch	0.5839	0.7101	0.7341	0.7174	0.7205	0.5785	0.7137	0.7441	0.6970	0.7183
(5) Ling 10ms	0.5823	0.7072	0.7391	0.7111	0.7191	0.5824	0.6912	0.7252	0.6930	0.7031
(6) Acous 50ms Ling 50ms	0.5411	0.7957	0.8354	0.8101	0.8137	0.5431	0.7801	0.8260	0.7883	0.7981
(7) Acous 10ms Ling 10ms	**0.5321**	**0.8194**	**0.8465**	**0.8141**	**0.8267**	**0.5358**	0.7992	0.8325	**0.7922**	**0.8080**
Two Subnets										
(10) Acous 50ms Ling 50ms	0.5420	0.7916	0.8303	0.8019	0.8079	0.5419	0.7804	0.8230	0.7873	0.7969
(11) Acous 10ms Ling 50ms	**0.5291**	**0.8323**	0.8526	**0.8236**	0.8362	**0.5311**	0.8188	0.8417	**0.8014**	**0.8206**
(12) Acous 50ms Ling Asynch	0.5416	0.7949	0.8385	0.7993	0.8109	0.5463	0.7685	0.8111	0.7724	0.7840
(13) Acous 10ms Ling Asynch	0.5296	0.8307	**0.8553**	0.8232	**0.8364**	0.5313	0.8203	**0.8437**	0.7935	0.8192
(14) Acous 10ms Ling 10ms	0.5310	0.8285	0.847	0.8189	0.8315	0.5316	**0.8205**	0.8381	0.7975	0.8187

Figure A.1: Results from Roddy et al., 2018b, and reproduction of those results. Reproductions reported here are averages of the results of models run by myself, Gruzin, 2020 and Smith, 2020; total 15 models trained for each no-subnets architecture, and 9 for two-subnets. Note: model numbers follow those used in Roddy's results, but I have not reproduced results for models with one subnet.

	Switchboard					Combined				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
No Subnets (Early Fusion)										
(1) Acous 50ms	0.4818	0.8290	0.8814	0.7916	0.8340	0.5235	0.7852	0.8295	0.7724	0.7957
(2) Acous 10ms	0.4717	0.8577	**0.8986**	0.7981	**0.8515**	0.5132	0.8213	0.8506	0.7939	0.8219
(3) Ling 50ms	0.5032	0.8191	0.8485	0.7692	0.8123	0.5517	0.7475	0.7781	0.7167	0.7474
(4) Ling Asynch	0.5088	0.8175	0.8444	0.7576	0.8065	0.5520	0.7485	0.7814	0.7105	0.7468
(5) Ling 10ms	0.5072	0.8159	0.8498	0.7711	0.8123	0.5566	0.7206	0.7582	0.7038	0.7275
(6) Acous 50ms Ling 50ms	0.4774	0.8445	0.8758	**0.8151**	0.8451	0.5172	0.8044	0.8480	0.8015	0.8178
(7) Acous 10ms Ling 10ms	**0.4705**	**0.8633**	0.8960	0.8031	0.8451	**0.5087**	**0.8285**	**0.8581**	**0.8047**	**0.8304**
Two Subnets										
(10) Acous 50ms Ling 50ms	0.4746	0.8480	0.8897	0.8277	0.8551	0.5179	0.8016	0.8417	0.8031	0.8155
(11) Acous 10ms Ling 50ms	**0.4653**	**0.8703**	**0.9046**	**0.8491**	**0.8747**	**0.5077**	**0.8361**	0.8607	**0.8137**	**0.8368**
(12) Acous 50ms Ling Asynch	0.4802	0.8373	0.8373	0.7980	0.8242	0.5212	0.7957	0.8331	0.7744	0.8011
(13) Acous 10ms Ling Asynch	0.4679	0.8670	0.9017	0.8414	0.8700	0.5089	0.8288	**0.8614**	0.8029	0.8310
(14) Acous 10ms Ling 10ms	0.4694	0.8665	0.9014	0.8204	0.8628	0.5088	0.8048	0.8592	0.8048	0.8229

Figure A.2: Results from reproducing experiments in Roddy et al., 2018b results, but for models trained on Switchboard, and models trained on Combined.

	Maptask (with *unk* tokens)				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
No subnets (early fusion)					
(1) Acous 50ms	0.5452	0.7650	0.8100	0.7784	0.7845
(2) Acous 10ms	0.5358	0.8074	0.8297	0.7958	0.8110
(3) Ling 50ms	0.5774	0.7132	0.7419	0.6895	0.7149
(4) Ling Asynch	0.5779	0.7134	0.7404	0.6817	0.7118
(5) Ling 10ms	0.5814	0.6893	0.7275	0.7021	0.7063
(6) Acous 50ms Ling 50ms	0.5417	0.7797	0.8306	0.7919	0.8007
(7) Acous 10ms Ling 10ms	0.5329	0.8037	0.8358	0.7970	0.8122
Two subnets					
(10) Acous 50ms Ling 50ms	0.5410	0.7841	0.8301	0.7842	0.7995
(11) Acous 10ms Ling 50ms	0.5295	0.8160	0.8398	**0.8042**	0.8200
(12) Acous 50ms Ling Asynch	0.5448	0.7699	0.8122	0.7817	0.7879
(13) Acous 10ms Ling Asynch	**0.5293**	**0.8240**	**0.8469**	0.7967	**0.8225**
(14) Acous 10ms Ling 10ms	0.5312	0.8196	0.8430	0.7968	0.8198

Figure A.3: Results of training different model configurations on Maptask data, but with low frequency words replaced with *unk* tokens, as discussed in Chapter 6. (1) and (2) were not retrained, since they do not use linguistic features.

	Switchboard (with *unk* tokens)					Combined (with *unk* tokens)				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
No Subnets (Early Fusion)										
(1) Acous 50ms	0.4818	0.8290	0.8814	0.7916	0.8340	0.5235	0.7852	0.8295	0.7724	0.7957
(2) Acous 10ms	**0.4717**	**0.8577**	**0.8986**	0.7981	**0.**8515	0.5132	0.8213	0.8506	0.7939	0.8219
(3) Ling 50ms	0.5070	0.8197	0.8472	0.7709	0.8126	0.5528	0.7485	0.7810	0.7073	0.7456
(4) Ling Asynch	0.5062	0.8199	0.8499	0.7655	0.8118	0.5522	0.7525	0.7783	0.7129	0.7479
(5) Ling 10ms	0.5076	0.8171	0.8497	0.7560	0.8076	0.5555	0.7203	0.7667	0.7034	0.7301
(6) Acous 50ms Ling 50ms	0.4778	0.8419	0.8781	**0.8119**	0.8440	0.5170	0.8052	0.8486	0.7955	0.8164
(7) Acous 10ms Ling 10ms	0.4726	0.8545	0.8916	0.8113	**0.8525**	**0.5070**	**0.8248**	**0.8568**	**0.8058**	**0.8291**
Two Subnets										
(10) Acous 50ms Ling 50ms	0.4748	0.8481	0.8881	0.8231	0.8531	0.5171	0.7980	0.8424	0.8010	0.8138
(11) Acous 10ms Ling 50ms	**0.4655**	**0.8690**	**0.9050**	**0.8507**	**0.8749**	**0.5061**	**0.8341**	0.8587	**0.8174**	**0.8367**
(12) Acous 50ms Ling Asynch	0.4800	0.8372	0.8823	0.8076	0.8424	0.5215	0.7924	0.8337	0.7730	0.7997
(13) Acous 10ms Ling Asynch	0.4684	0.8659	0.9007	0.8464	0.8710	0.5077	0.8335	0.8606	0.8090	0.8344
(14) Acous 10ms Ling 10ms	0.4696	0.8633	0.9002	0.8287	0.8641	0.5091	0.8290	**0.8625**	0.8113	0.8343

Figure A.4: Results of training different model configurations on Switchboard and the combined data set, but with low frequency words replaced with *unk* tokens, as discussed in Chapter 6. (1) and (2) were not retrained, since they do not use linguistic features.

	Train on Maptask					Train on Switchboard				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
Test on Maptask										
(A) Acous 10ms	0.5358	0.8074	0.8297	0.7923	0.8098	0.6039	0.6943	0.7271	0.6800	0.7005
(B) Ling 50ms	0.5794	0.7081	0.7325	0.6646	0.7017	0.6350	0.6445	0.5732	0.5834	0.6004
(C) Acous 10ms Ling 50ms	0.5291	0.8214	0.8487	0.8009	0.8237	0.6001	0.7071	0.7394	0.7020	0.7162
Test on Switchboard										
(A) Acous 10ms	0.6179	0.7282	0.7415	0.6936	0.7211	0.4717	0.8577	0.8986	0.7981	0.8515
(B) Ling 50ms	0.5938	0.7689	0.8209	0.6121	0.7340	0.5032	0.8191	0.8485	0.7692	0.8123
(C) Acous 10ms Ling 50ms	0.5842	0.7686	0.7798	0.7479	0.7654	0.4653	0.8703	0.9046	0.8491	0.8747
Test on Combined										
(A) Acous 10ms	0.5657	0.7758	0.7900	0.7538	0.7732	0.5547	0.7491	0.7837	0.7293	0.7540
(B) Ling 50ms	0.5836	0.7302	0.7539	0.6475	0.7105	0.5874	0.7032	0.6637	0.6593	0.6754
(C) Acous 10ms Ling 50ms	0.5493	0.7963	0.8117	0.7782	0.7954	0.5495	0.7618	0.7946	0.7604	0.7723

Figure A.5: Results for cross-corpus tests on models trained on Maptask, and models trained on Switchboard, as described in Chapter 6

	Train on Combined				
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**
Test on Maptask					
(A) Acous 10ms	0.5373	0.7960	0.8243	0.7860	0.8021
(B) Ling 50ms	0.6680	0.5517	0.5464	0.6540	0.5840
(C) Acous 10ms Ling 50ms	0.6167	0.5228	0.5693	0.7894	0.6272
Test on Switchboard					
(A) Acous 10ms	0.4724	0.8553	0.8993	0.8037	0.8528
(B) Ling 50ms	0.5118	0.8210	0.8565	0.7954	0.8243
(C) Acous 10ms Ling 50ms	0.4706	0.7637	0.8405	0.8757	0.8266
Test on Combined					
(A) Acous 10ms	0.5132	0.8213	0.8506	0.7939	0.8219
(B) Ling 50ms	0.5517	0.7475	0.7781	0.7167	0.7474
(C) Acous 10ms Ling 50ms	0.5077	0.8361	0.8607	0.8137	0.8368

Figure A.6: Results for cross-corpus tests on models trained on Combined, as described in Chapter 6

	Test on both Maptask roles					Test on F		Test on G	
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**F1 onset**	**BCE Loss**	**F1 onset**
Train on both roles									
(A) Acous 10ms	0.5358	0.8074	0.8297	0.7958	0.8110	0.4526	0.7816	0.6189	0.7670
(B) Ling 50ms	0.5774	0.7132	0.7419	0.6895	0.7149	0.4904	0.5871	0.6644	0.6800
(C) Acous 10ms Ling 50ms	0.5295	0.8160	0.8398	0.8042	0.8200	0.4439	0.7821	0.6149	0.7805
Train on F									
(A) Acous 10ms	0.5616	0.7579	0.8150	0.7698	0.7809	0.4550	0.7935	0.6685	0.7298
(B) Ling 50ms	0.6178	0.6603	0.6820	0.6336	0.6586	0.4904	0.6789	0.7450	0.5993
(C) Acous 10ms Ling 50ms	0.5553	0.7715	0.8279	0.7774	0.7923	0.4479	0.8043	0.6632	0.7333
Train on G									
(A) Acous 10ms	0.5583	0.7716	0.8130	0.7400	0.7749	0.4954	0.7236	0.6213	0.7707
(B) Ling 50ms	0.6222	0.6591	0.6554	0.5898	0.6348	0.5783	0.6136	0.6660	0.5803
(C) Acous 10ms Ling 50ms	0.5516	0.7915	0.8318	0.7633	0.7955	0.4860	0.7672	0.6174	0.7588

Figure A.7: Results of role experiments on Maptask (Chapter 7) investigating difference in model performance on data from *instruction followers* (F), and *instruction givers* (G)

.

	Test on both Switchboard roles					Test on A		Test on B	
	BCE Loss	**f1 50ms**	**f1 500ms**	**f1 onset**	**Avg. F1**	**BCE Loss**	**F1 onset**	**BCE Loss**	**F1 onset**
Train on both roles									
(A) Acous 10ms	0.4717	0.8577	0.8986	0.7981	0.8515	0.4813	0.8192	0.4649	0.7913
(B) Ling 50ms	0.5070	0.8197	0.8472	0.7709	0.8126	0.5134	0.8219	0.5006	0.7081
(C) Acous 10ms Ling 50ms	0.4655	0.8690	0.9050	0.8507	0.8749	0;4731	0.8819	0.4579	0.8178
Train on A									
(A) Acous 10ms	0.5705	0.7340	0.7682	0.7194	0.7405	0.4920	0.8085	0.4765	0.7602
(B) Ling 50ms	0.5111	0.8180	0.8430	0.7609	0.8073	0.5172	0.8014	0.5049	0.7164
(C) Acous 10ms Ling 50ms	0.4760	0.8570	0.8964	0.8293	0.8609	0.4854	0.8692	0.4668	0.7847
Train on B									
(A) Acous 10ms	0.5634	0.7387	0.7712	0.7257	0.7452	0.4864	0.8208	0.4714	0.7625
(B) Ling 50ms	0.5120	0.8167	0.8456	0.7478	0.8034	0.5189	0.7999	0.5051	0.6907
(C) Acous 10ms Ling 50ms	0.4756	0.8616	0.8972	0.8266	0.8618	0.4848	0.8654	0.4665	0.7840

Figure A.8: Results of role experiments on Switchboard (Chapter 7) investigating difference in model performance on data from roles *A* and *B*. These are not defined roles in Switchboard, other than the fact that participant *A* initiates the call; both speakers are given the same conversation prompt (Godfrey and Holliman, 1993.)

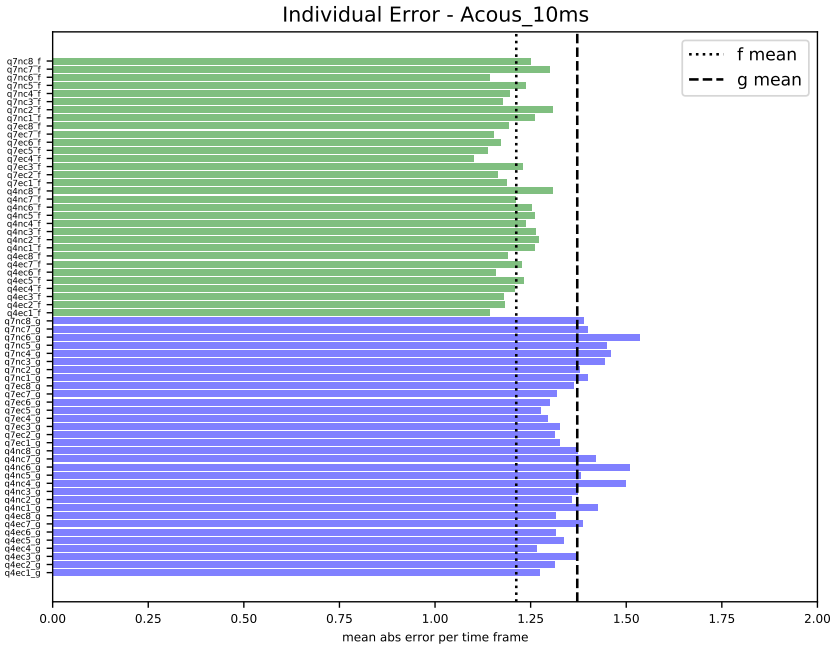

Figure A.9: MAE per person for acoustic-only models on Maptask. F is in green, G is in blue.

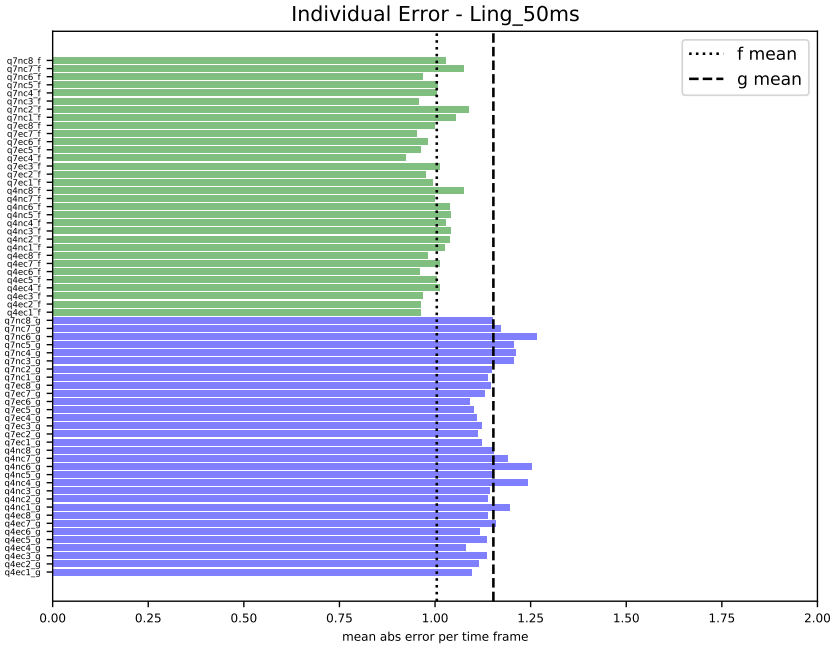

Figure A.10: MAE per person for linguistic-only models on Maptask. F is in green, G is in blue.

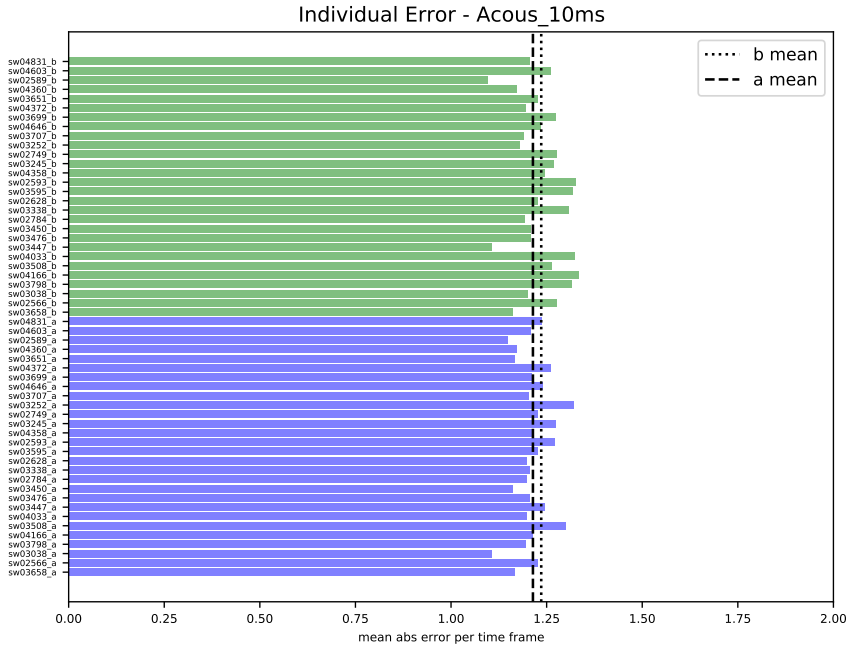

Figure A.11: MAE per person for acoustic-only models on Maptask. B is in green, A is in blue.

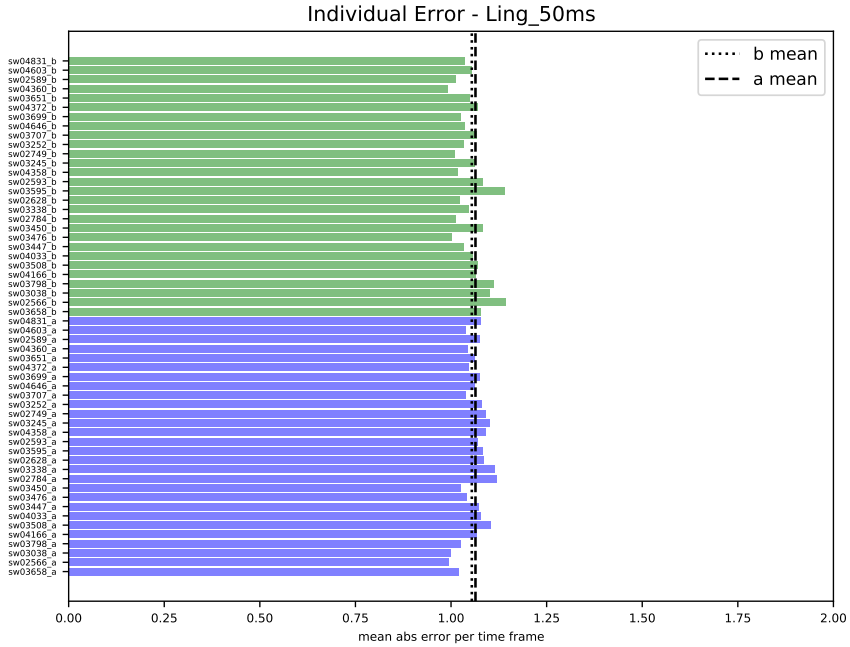

Figure A.12: MAE per person for linguistic-only models on Maptask. B is in green, A is in blue.