Accent-Emotion Entanglement in LM-Based Text-to-Speech Systems

Matt Hayden, Jinzuomu Zhong,*

*Korin Richmond**

**Supervisors*

Master of Science

Speech and Language Processing

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2025

Abstract

Many zero-shot TTS systems now claim to produce near-human quality imitations of speech. However, the recent growth in features and the rush to report state-of-the-art results can lead to unexpected issues slipping through the cracks. This paper investigates such an issue, a new phenomenon dubbed accent-emotion entanglement. It is found that, in CosyVoice 2 [Du et al., 2024b], an instruction-based zero-shot TTS system, accent hallucination is caused and guided by the emotion of the input provided to the model. A similar zero-shot model, MaskGCT [Wang et al., 2024], does not produce the same results. The paper also discusses the importance of subjective results in evaluation as a means to avoiding future unexpected issues.

Acknowledgements

First and foremost I would like to thank Jinzuomu Zhong. His supervision has been truly invaluable throughout. There were too many instances to count where I felt stuck or lost and without fail he came to my rescue. I've learned so much from working with him and hope to continue that work in the future as we continue this project. I would also like to thank Dr. Korin Richmond for his help over the last 3 months. Meetings every Wednesday always seemed to bring up new questions, new solutions, new challenges and new reassurance. I could not have pivoted so easily from one project to another without his support and the occasional chat during a coffee morning. I must also give my thanks to the Speech and Language Processing MSc group as a whole. Prof. Simon King and Dr. Catherine Lai have been so kind and supportive across all 3 semesters guiding and giving me the confidence to succeed in a whole new venture. Finally, thank you to all my loved ones who have never wavered in their support. To my parents, who started me on my journey, taught me to find joy in learning and gave me the perseverance to continue when the going gets tough, as it often does. To my love, for bringing me respite and joy (and pastries) throughout such a trying period. And to my friends, you know who you are, who provided an ear to complain to, a pub quiz to attend and a laugh when it was needed most. Thank you.

Table of Contents

Chapter 1

Introduction

1.1 Motivation: Uncontrolled and Under-Evaluated TTS Systems

In the past few years, much focus in TTS research has shifted to zero-shot systems. Zero-shot TTS systems focus on imitating a previously unseen speaker given only a short audio clip of the individual. Many zero-shot TTS system now impressively claim to produce speech close or equal in quality to human recordings [Casanova et al., 2024, Jiang et al., 2023, Jiang et al., 2024]. However, deeper analysis of these results reveals that the evaluation metrics employed in these papers are often either poorly described, overburdened or underutilised. Imitation quality is often measured using only objective speaker similarity scores. It can be unclear what such metrics capture or how exactly they measure perceptual differences in speech such as timbre, accent and emotion. At the same time, the number of features offered by such models, as well as the amount of data used to train them, is ever increasing. With such large datasets and so many possible configurations, it can be hard to anticipate, measure or control the numerous unforeseen errors that may occur. All these issues can be exacerbated if aspects of the TTS system, such as evaluation metrics, data preprocessing or training data makeup, are poorly described or not made publicly available. This leaves no opportunity for other researchers to discover, replicate or solve unexpected issues.

While investigating emotion controllability in CosyVoice 2 [Du et al., 2024b], a zero-shot TTS system which accepts generation guidance through natural language instructions, I discovered one such unexpected issue. This testing required that CosyVoice

2 be provided with emotional input. Emotional input here is defined as: an instruction requesting the synthesised speech be given a certain emotion as well as reference speech where the speaker matches this emotion. For example, the instruction may be 'Make this sound sad.' and the reference speech is that of a sad speaker from an emotional speech dataset. If an instruction cannot be given due to the nature of the system, emotional input refers exclusively to the emotion displayed by the reference speaker. When prompting CosyVoice 2 with emotional input, the speaker in the synthesised utterances had an unexpected accent, one which greatly differed from that of the human reference speaker. This constitutes accent hallucination, a phenomenon which occurs when input to a TTS system uses one accent while the synthesised output uses either a completely different accent or an accent that is a blend of the original and 1 or more novel accents. The specific accent hallucinated by CosyVoice 2 appeared to be impacted by the emotion of the input. Angry emotional input produced differing results to sad emotional input. Importantly, neutral input, that is reference speech with no obvious emotion and no instruction given, did not seem to lead to accent hallucination.

This was a concerning finding. Accents are integral not only to understanding but also to identity. Removing a speakers accent or switching it for another is unlikely to be well received. However, while any form of accent hallucination is an issue, accent hallucination tied to emotion is especially problematic as it may lead the TTS system propagating harmful stereotypes which link emotions to certain cultures or people, such as that of the 'angry black woman' [Walley-Jean, 2009]. The seriousness of this issue meant it required further investigation so that we may better understand its causes and thus begin to propose solutions.

1.2 Research Gap: The Links Between Accent and Emotion

To the best of my knowledge, I am the first to both report and subsequently investigate this phenomenon of emotional input leading to accent hallucination. I have dubbed it accent-emotion entanglement. Some research [Zhu et al., 2024, Xin et al., 2021, Badlani et al., 2023] has explored the disentanglement of accent information from speaker embeddings in order to produce speech without the accent of the original speaker. The goal of such systems, however is generally to split a speaker's accent from the other characteristics

that define their voice so that their original accent may be swapped for a novel one. Accent hallucination in these cases would refer to leaking of the speaker's original accent into the synthesised speech. This differs from the case observed in CosyVoice 2 where the hallucinated accent is not present anywhere in the input, not in the reference speech nor in the instruction. Accent-emotion entanglement is a novel, unexplored phenomenon.

Chapter 2

Background

The first step in investigating accent-emotion entanglement is to confirm the issue actually exists. Having been discovered initially by accident, it is important to establish how widespread and systematic the accent hallucination may be and, beyond this, determine if the root cause is indeed emotional input. This section introduces and briefly discusses the results of a listening test carried out with the goal of studying the prevalence of accent-emotion entanglement in TTS systems as well as the factors which contribute to the accent hallucination.

2.1 The TTS Systems: CosyVoice 2

The focus of this paper is on the CosyVoice 2 [Du et al., 2024b] TTS system, specifically on its natural language instruction mode. CosyVoice 2 was initially chosen due to claims of high performance, open-source availability and its feature of accepting both an instruction and a reference speaker. The model is based around the Qwen2.5-0.5B textual LLM. All input into CosyVoice 2 is tokenised and this, along with the language model, allows it to treat TTS as a next token prediction problem. The input in CosyVoice 2 is threefold. The text to be synthesised, henceforth target text, as well as the transcription of the reference utterance, is tokenised using Byte-Pair Encoding. The tokens are then passed through the Qwen embedding layer. This same process is applied to the instruction given to CosyVoice 2. These instructions are quite simple in structure allowing only simple tags or directions. These may include the desired emotion, speaking rate and, for synthesis in Chinese, certain dialects. Fine-grained instructions, vocal bursts or features such as laughter or breaths, may also be included. This differs from other models which accept natural language instructions,

such as Parler-TTS [Lyth and King, 2024], which allows much more freedom in the style of instructions. Finally, CosyVoice 2 requires reference audio, a short recording of an individual whose voice will be replicated in the synthesised utterance. This reference audio is processed using a supervised tokeniser, which projects windows of speech down into low rank representations. These are then input as tokens into the text-speech LLM. Tokens produced by this LLM are said to focus on semantic information only, these tokens are later passed to a conditional flow matching model which predicts the acoustic information, such as timbre, and produces mel spectrogram frames which are finally passed to a pre-trained vocoder for synthesis. CosyVoice 2 has 4 possible configurations. The realisation of the synthesised speech can be controlled via 2 different modes. In the in-context learning mode, the system is guided only by reference audio while in the instruction mode, it is led by both reference audio and a natural language instruction. Additionally, streaming and non-streaming modes are available. In streaming mode speech is produced as tokens are predicted. In non-streaming mode, speech is only produced after all tokens have been predicted.

The training data used for CosyVoice 2 is largely described only in vague terms. Although CosyVoice 2 is capable of synthesis in 4 different languages, principally Chinese, the focus of this paper is only on synthesis of English using English instructions and English reference speech. The model is trained on 30,000 hours of data for English from an in-house dataset, the details of which are not revealed. However, processing of this data is performed automatically using SenseVoice [An et al., 2024]. This processing includes ASR for producing transcriptions, audio-event detection and utterance-level emotion recognition. Notably, it does not label accent in any way. An additional 1,500 hours of natural language and fine-grained instruction training data is used for fine-tuning. There are no details on the make-up of this dataset. The lack of information regarding the data used in CosyVoice 2 makes it difficult to be pinpoint the exact cause of any accent hallucination. This will be explored further in Chapter 6.

2.2 The TTS Systems: MaskGCT

In order to investigate the pervasiveness of accent hallucination in emotional TTS, a second TTS system is employed in all experiments. MaskGCT [Wang et al., 2024] was chosen for this due to both its similarities and differences to CosyVoice 2. Both models use a two-stage system, converting input first to semantic tokens, then to acoustic ones.

However, where the database used in CosyVoice 2 is unclear, MaskGCT uses a publicly available in-the-wild dataset, Emilia [He et al., 2025]. MaskGCT is trained using 50,000 hours of English data from the Emilia. The dataset is built from in-the-wild data taken from online sources such as YouTube and podcasts and is both acoustically and semantically more diverse than typical datasets. It is not labelled with information on accent or emotion. MaskGCT is trained by predicting masked tokens, a self-supervised task that doesn't require explicit labels.

CosyVoice 2 and MaskGCT also offer different approaches to controllability in TTS. CosyVoice 2 accepts instructions whereas MaskGCT is fully focused on in-context learning, taking input exclusively from the target text and reference speech. Both inputs are encoded and tokenised then used to produce semantic tokens which in turn are used to predict acoustic tokens which are finally used to predict mel spectrogram. These are then passed to a vocoder. MaskGCT does not accept instructions, it extracts any emotional or accent information solely from the reference speech provided. The similarities and differences between MaskGCT and CosyVoice 2 may give insight into the root cause of accent-emotion entanglement. This will be explored further in Section 5.

2.3 Results Claims: CosyVoice 2 vs. MaskGCT

Direct comparison of any 2 TTS systems can be difficult given that reported results often use differing test sets, rebuilt models and a wide variety of evaluation metrics. However, to get an idea of what to expect from the experiments in this paper it is important to consider pre-existing evaluations. Table 2.1 shows the results reported for CosyVoice 2 along with MaskGCT results for comparison. It should be noted that these results are from CosyVoice 2's in-context learning mode. Instruction mode results are only available for Chinese speech. Speaker similarity is measured using both Eres2Net [Chen et al., 2023], in brackets, and a WavLM-based metric and testing is performed using 2 datasets taken from SEED-tts [Anastassiou et al., 2024], both in English. It is not clear how these speaker similarity measures account for a speaker's accent. What is clear from the results is that there is a lot of variability between systems and metrics. Using the WavLM-based metric, MaskGCT outperforms both CosyVoice and CosyVoice 2 across both test sets. However, using the ERes2Net speaker similarity, CosyVoice 2 prevails with MaskGCT also outperformed by the original CosyVoice.

Model	*test-en*	*test-hard*
	SS ↑	SS↑
Human	0.734 (0.742)	-
Vocoder Resyn.	0.700	-
MaskGCT	**0.714 (0.730)**	**0.748** (0.720)
CosyVoice	0.609 (0.699)	0.709 (0.755)
CosyVoice 2	0.652 (0.736)	0.724 **(0.776)**

Table 2.1: Speaker similarity scores derived from a WavLM-based model and ERes2Net (in brackets). Results taken from Du et al. 2024b. Bold represents highest score for that test set and similarity metric.

Eres2Net is used throughout the rest of the CosyVoice 2 paper. It is difficult to glean insight into future performance from these results due to their variability. However, they do point to a broader issue in the use of objective evaluation metrics. It is not completely clear what either speaker similarity metric is capturing. Human speech is multi-dimensional and what defines a speaker is far from certain. Given this, it is perhaps unsurprising that catch-all metrics such as speaker similarity can show so much variety.

Results given in Wang et al (2024), shown in Table 2.2 do offer more clarity. Unfortunately, CosyVoice 2 is not included in comparisons, only its predecessor [Du et al., 2024a]. However, the greater variety of evaluation metrics still helps to paint a better picture of expected performance. SIM-O represents overall speaker similarity, calculated using cosine similarity. The method used to create embeddings for this metric is not given. The paper notes however, that this metric is limited and thus emotion and accent are also explored in order to give a better picture of speaker imitation quality. Accent SIM and Emotion SIM also represent cosine similarity, however the embeddings for these metrics are derived from the CommonAccent [Zuluaga-Gomez et al., 2023] accent recognition system and emotion2vec [Ma et al., 2023] an emotion recognition system, respectively. Importantly, SMOS is also reported for both accent and emotion. Subjective results help to ground their objective counterparts in human perception. For accent metrics, the ground truth is taken from the L2-ARCTIC [Zhao et al., 2018] dataset while the emotional ground truth is from the Emotional Speech Dataset [Zhou et al., 2021].

The results given show that in terms of accent, MaskGCT comfortably outperforms

System	Accent			Emotion		
	SIM-O ↑	**Cos Sim ↑**	**SMOS ↑**	**SIM-O ↑**	**Cos Sim ↑**	**SMOS ↑**
Ground Truth	0.747	0.633	-	0.637	0.936	-
CosyVoice	0.653	0.640	3.99 $\pm$ 0.23	0.575	0.839	3.66 $\pm$ 0.19
MaskGCT	0.717	0.645	4.38 $\pm$ 0.25	0.600	0.822	3.76 $\pm$ 0.25

Table 2.2: Evaluation results for accent and emotion imitation. Taken from Wang et al. (2024).

CosyVoice. The emotion scores are much closer with a mismatch between objective and subjective results. Given we may expect to see some improvement from CosyVoice to CosyVoice 2, this indicates that CosyVoice 2 and MaskGCT may perform similarly in terms of emotion, however the accent score gap may be more difficult to bridge. These results once again show the variance that can follow from the use of objective metrics, reaffirming the importance of subjective results alongside them. The speaker similarity score here align with the SMOS scores but not with the ERes2Net speaker similarity scores in Table 2.1 or with the emotional cosine similarities. This points to the need for subjective results which can help ground further objective metrics when establishing the prevalence of accent-emotion entanglement.

Chapter 3

Listening Test Design

Having explored both CosyVoice 2 and MaskGCT, this chapter focuses on the listening test used to investigate accent-emotion entanglement in the 2 systems. It should be noted that for CosyVoice 2, all utterances were synthesised using the instruction, non-streaming mode. When referencing utterance configurations throughout this paper, a consistent naming scheme is used: *system-target_text-emotion*. Systems are abbreviated: CosyVoice 2 = *CV2*, MaskGCT = *MGCT* and ground truth = *GT*. Thus, *CV2-021-Happy* refers to an utterance synthesised with CosyVoice 2, target text *021* and using *happy* emotional input.

3.1 The MEAD Dataset

The target text, reference sentences and ground truth utterances for this listening test were all taken from the MEAD dataset [Wang et al., 2020]. MEAD is an audio-visual database consisting of 60 actors producing utterances with 8 different emotions at 3 intensity levels. The video component of MEAD is entirely ignored in this paper. There are 13 unique sentences that are shared across all emotions, all of which are themselves neutral in sentiment. Each utterance is recorded only once per intensity level.

3.2 Speaker Selection

The same constraints also meant that only one speaker could be studied. Further speakers will be explored using objective metrics. During initial exploration of the TTS systems, it was noted that CosyVoice 2 almost never replicated the accent of the reference

speaker. Instead the accent defaulted to General American (GenAm). This occurred regardless of reference speaker, instruction or target text and therefore appears distinct from the hallucination caused by emotional input. The aim of this paper is to investigate accent-emotion entanglement exclusively. Study of other forms of accent hallucination, such as giving speakers a GenAm accent regardless of their true accent, is left for future work. The MEAD database give very little meta information on readers, only their gender. Given this, and the need for a GenAm speakerm a brief study was conducted to find the MEAD speaker whose accent was closest to GenAm. Using a GenAm speaker avoids CosyVoice 2 generating a GenAm accent from being treated as accent hallucination by listening test participants. The study involved 5 expert listeners ranking 5 MEAD speakers in terms of accent similarity to a GenAm speaker, taken from the CommonAccent dataset [Zuluaga-Gomez et al., 2023]. The gender of the MEAD speakers differed from that of the reference. 10 male speakers and 10 female speakers were chosen for this study at random. The highest scoring speaker, thus the one with an accent most similar to GenAm, was M007, a male speaker. This is the speaker used in the listening test.

3.3 Reference and Target Text Selection

3.3.1 Split into References and Target Text

Limited budget again meant a constraint on the number of audio clips that could be presented in the listening test. Thus, to ensure those chosen were maximally informative, both the target texts and reference sentences were carefully chosen. The 13 sentences shared between all emotions are split by MEAD into 3 common sentences and 10 generic sentences. All sentences are given in Appendix A. The 3 common sentences were chosen to act as reference sentences, with the 10 generic sentences used as target text. Time constraints meant that testing could only be carried out using 2 reference sentences. Sentence *003* was excluded as it was markedly shorted than sentences *001* and *002*. Sentence *025* was removed from the set of target texts due to concerns its content was in some way connected to accent through race. There is no guidance on the style of instructions given to CosyVoice 2. They were thus designed to be a simplistic as possible to avoid confusion. Neutral utterances received no instruction, angry utterances 'Make this sound angry.' and happy utterances 'Make this sound happy'. Further work could explore the impact of instruction makeup.

3.3.2 Quantifying the Effects of Reference Sentence and Target Text

The goal of systematic reference and target text selection is to uncover trends that can be further explored through the listening test. To do this, each of the 9 target texts was synthesised 100 times using each of the reference sentences, for each emotion using each system. This resulted in 10,800 total synthesised utterances. It was important to synthesise each combination 100 times as CosyVoice 2 is a probabilistic model liable to inconsistency and earlier testing had shown that accent hallucination was not consistent even when every condition was kept the same. To capture the variability in accent, each utterance was passed to an accent recognition software, GenAID [Zhong et al., 2025], to produce a probability distribution across 13 accents (Given in Appendix B). This distribution was then averaged across the 100 utterances for each combination. Finally, the entropy of each averaged distribution was calculated in order to measure how GenAID's certainty changed across conditions. This was calculated as:

$$\sum_{i=0}^{k} p(x_i) \cdot ln(p(x_i))$$

where $p(x_i)$ represents the probability given to an accent an $k = 13$. A low entropy represents GenAID on average being certain about the accent of a combination, higher entropy may point to accent hallucination occurring. This metric was chosen as we are not interested in single categorical labels. Instead we wish to see how accent broadly shifts across multiple generations. Table 3.1 gives the absolute difference in entropy between reference sentences *001* and *002* for each system, when averaged across all target texts. The effect of reference sentence appears to be minimal for both systems. This makes sense since semantic information is not extracted from the text of the reference utterance in either. For clarity and ease of presentation, the remaining entropy results in the section are the average results of reference sentence *001* and *002*.

Table **??** gives the entropy results for each target text. From this, we can see that CosyVoice 2 seems to produce utterances with the highest entropy across nearly target texts for all 3 emotions suggesting variability in the accents it produces. Meanwhile, the entropy scores for MaskGCT tend to be close to those for the equivalent ground truth utterances. This shows that, MaskGCT is consistent in the accents it produces and perhaps that it is imitating the ground truth closely.

Despite the consistent ranking of systems, these results do point to the the target

Emotion	CosyVoice 2	MaskGCT
Angry	0.002	0.014
Happy	0.013	0.010
Neutral	0.025	0.010

Table 3.1: Difference in entropy, averaged across target texts, between utterances using reference sentence *001* and those using reference sentence *002*.

text having an impact of accent realisation in CosyVoice 2. *CV2-021* has high entropy across the board but still shows an increase for *angry* and *happy* utterances compared to *neutral* ones. For *CV2-022* on the other hand, *neutral* utterances have the greatest entropy. *CV2-023* has the lowest entropy across nearly all systems and for CosyVoice 2 sees a rise in entropy from *neutral* to *happy* but a fall from *neutral* to *angry*. *CV2-026* has the greatest difference between both *happy* and *angry* utterances and *neutral* ones for CosyVoice 2, suggesting it may suffer the most from accent-emotion entanglement. For MaskGCT and the ground truth, there are no such shifts. Finally, for target text *027* MaskGCT's entropy is close to or even greater than CosyVoice 2. These 5 target texts are the ones used in the listening test. The variation in entropy scores will compared to listening test results, allowing for validation of this objective metric. Finally, it should be noted that the extremely low entropy for the *GT-021-Neutral* appears to stem from an issue with GenAID. The system is 97% sure the speaker is Australian when this is not the case.

3.4 Utterance Selection

Each condition, i.e. system, emotion and target text, will only appear twice in the listening test, once to measure accent similarity, once to measure emotion similarity. It is thus important to ensure the utterances used can be maximally informative. To achieve this, following Wang et al (2024), the cosine similarity between each utterance and its equivalent ground truth utterance was calculated. The embeddings for each utterance were extracted using GenAID [Zhong et al., 2025]. Then, the utterance with the lowest cosine similarity was chosen for each condition. This was done to give the best chance of finding utterances with accent hallucination. Utterances could be produced using either reference sentence. For MaskGCT some utterances had to be

Emotion	System	Target Text								
		021	022	023	024	026	027	028	029	030
Angry	Ground Truth	0.71	**0.81**	0.71	0.70	0.67	0.68	20.70	0.70	*0.67*
	CosyVoice 2	1.33	0.98	*0.63*	0.89	**1.35**	0.69	1.33	1.23	0.91
	MaskGCT	0.94	**1.13**	*0.70*	0.73	0.74	0.86	0.72	0.72	0.76
Happy	Ground Truth	**0.98**	0.73	*0.68*	0.70	0.70	0.69	0.70	0.71	0.70
	CosyVoice 2	**1.73**	0.87	0.88	1.18	1.56	*0.72*	1.22	1.02	1.08
	MaskGCT	0.79	**0.83**	*0.69*	0.72	0.71	0.70	0.70	0.71	0.71
Neutral	Ground Truth	*0.15*	**0.72**	0.71	0.71	0.72	0.70	0.67	0.69	0.68
	CosyVoice 2	**1.18**	1.14	*0.71*	1.03	0.82	0.77	0.78	0.78	0.85
	MaskGCT	**1.03**	0.94	*0.70*	0.77	0.72	0.71	0.71	0.71	0.73

Table 3.2: Average entropy for each target text across each emotion for each system. Green indicates lowest entropy system for respective target text and emotion, red the highest. Bold indicates entropy target text for a system across that emotion, italics the lowest.

excluded as they included artifacts which would make accent or emotion judgements difficult. In such cases, the utterance with the next lowest cosine similarity and no artifacts was chosen.

3.5 Test Design

The goal of the listening test was to measure the similarity between a synthesised utterance and its ground truth equivalent in terms of accent and emotion. Each question asked about only one of accent or emotion similarity to avoid one judgement affecting the other. This resulted in 60 questions total (+5 attention checks). Participants were asked to rate similarity from 1 (no similarity) to 5 (nearly identical). Example questions are given in Appendix C.

30 participants were recruited using Prolific[1]. They were balanced for gender and ranged in age from 25-64. Test takers were also evenly spread across the United Kingdom, United States, Australia, Ireland and South Africa. This was done to increase the

[1]https://www.prolific.com/

diversity in accent of the test takers themselves. All participants had English as the L1 and primary language.

3.6 Hypotheses

The results of the listening test are used to explore 2 key questions. Is emotional input causing accent hallucination in CosyVoice 2? If that's the case, is the phenomenon also present in MaskGCT. Beyond accent, the test will also give information on how accent hallucination impacts emotion realisation in CosyVoice 2. Initial brief testing found accent hallucination in *angry* and *happy* utterances, but never in *neutral*. This was supported by the entropy results which suggested that, at least for some target texts, there was much less certainty about accent in emotional utterances than in neutral ones. These results also clearly demonstrated that no such uncertainty occurs with MaskGCT. This supports the results presented in Wang et al. (2024), which showed MaskGCT outperforming CosyVoice in all accent metrics. Results in that same paper also suggested that performance on emotion realisation may be similar between the 2 systems. Given this, the hypotheses for the listening test are threefold:

- For CosyVoice 2, all emotional utterances (*happy*/*angry*) will receive lower accent SMOS scores than their neutral *equivalents*.
- All MaskGCT emotional utterances will outperform their CosyVoice 2 counterparts in accent SMOS but there will not be a significant difference between *neutral* utterances.
- There will be no significant difference between CosyVoice 2 and MaskGCT utterances in terms of emotion SMOS.

Chapter 4

Results

The listening test was carried out without issue. 1 participant was excluded for failing attention checks. The results of the remaining 29 test takers are presented in this chapter. After this, SMOS scores are compared to objective measures to explore aspects of accent-emotion entanglement that could not be covered in the subjective evaluation. A selection of the utterances discussed in this section can be heard online[3].

4.1 Subjective Accent Results

Table 4.1 presents the accent SMOS scores from the listening test. The initial stand out results come from CosyVoice 2 with extremely low scores (¡2) across multiple positions. These can be clearly seen in Figure **??**. This suggests that CosyVoice 2 is producing utterances with hallucinated accents. In addition to this, the results for CosyVoice 2 vary much more than those of MaskGCT, with standard deviations across all scores of 1.16 vs. 0.26. This aligns with the entropy results which showed MaskGCT to be very consistent whereas CosyVoice 2 is more unpredictable. Not all utterances see accent hallucination. It also seems as if the target text has an effect on accent for CosyVoice 2, but not for MaskGCT. For example, for *CV2-022* and *CV2-023* SMOS is high for *angry* but very low for *happy* and *neutral* utterances. *CV2-021* and *CV2-026* however see lower scores for the angry and happy utterances than the neutral. These patterns point to a larger finding, some neutral utterances appear to also see accent hallucination. Listening to the low scoring *neutral* utterances confirms this. For *CV2-022-Neutral*, *CV2-023-Neutral* and *CV2-026-Neutral* an obviously British accent can be heard. For *CV2-021-Neutral* and *CV2-027-Neutral*, the accent is GenAm. How this accent hal-

[3]https://matthayden23.github.io/

System	Emotion	Accent SMOS Scores by Target Text				
		021	**022**	**023**	**026**	**027**
CosyVoice 2	**Angry**	1.33 ± 1.03	**3.54 ± 1.2**	**3.04 ± 1.35**	*1.25 ± 0.86*	*2.63 ± 1.24*
	Happy	*1.17 ± 0.49*	*1.42 ± 0.79*	*1.21 ± 0.85*	*1.25 ± 0.85*	**4.13 ± 1.06**
	Neutral	**4.13 ± 0.87**	1.67 ± 1.03	1.75 ± 1.01	**2.04 ± 1.35**	4.00 ± 0.85
MaskGCT	**Angry**	*3.75 ± 1.14*	*3.33 ± 1.18*	*3.75 ± 1.01*	**3.92 ± 1.2**	**4.13 ± 0.9**
	Happy	3.79 ± 1.2	**4.04 ± 1.15**	3.79 ± 0.8	3.5 ± 1.04	*3.83 ± 1.07*
	Neutral	**4.33 ± 0.78**	3.96 ± 0.71	**3.83 ± 1.07**	*3.42 ± 0.78*	*3.83 ± 1.03*

Table 4.1: Accent SMOS and standard deviation results from listening test. Bold represents highest SMOS for that system and target text, italics lowest SMOS.

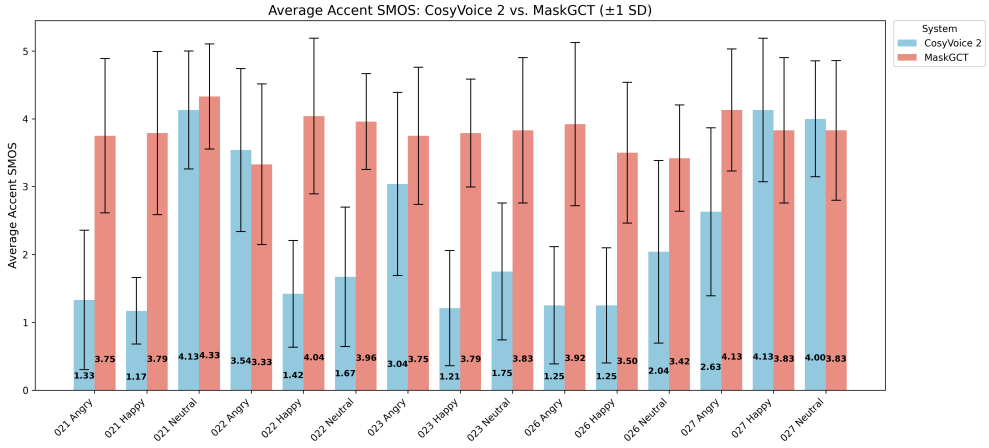


Figure 4.1: Average accent SMOS for CosyVoice 2 and MaskGCT following listening test with 29 English-speaking participants.

lucination is realised will be explored further in Section 4.2, however these findings falsify the hypothesis that *neutral* utterances will outperform their *angry* and *happy* counterparts. It seems that accent hallucination is not exclusively caused by emotional input.

It is though noticeable that happy utterances for CosyVoice 2 have particularly low scores. Where they are close in score to their neutral equivalents this is because the neutral score is also low. This is not true for MaskGCT where the differences are minimal. There is little variation across MaskGCT scores for all emotions with utterances generally scoring well. Table 4.2 gives the average SMOS across target texts for each emotion, as well as the p-values obtained from an unpaired t-test.

System	Average Accent SMOS by Emotion		
	Angry	**Happy**	**Neutral**
CosyVoice 2	2.36 ± 1.03	1.83 ± 1.29	2.72 ± 1.24
MaskGCT	3.78 ± 0.29	3.79 ± 0.19	3.88 ± 0.33
p-value	**0.018**	**0.0099**	0.078

Table 4.2: Accent SMOS scores averaged across target texts and p-values comparing the systems. Bold represents a significant result ($p < 0.05$).

These results are line with the predictions made previously. There is a significant difference between CosyVoice 2 and MaskGCT for the *angry* and *happy* scores but not for *neutral*. However, the difference between the 2 systems for *neutral* utterance accent SMOS is still quite large due to the accent hallucination in *CV2-022*, *-023* and *-026*. Finally, the standard deviation across scorers, given in Table 4.1 while consistent, is large. This demonstrates the difficulty in evaluating accents, even for humans. Much of this difficulty can come from separating a speaker's accent from the other characteristics of their speech, such as emotion or prosody. This is equally and issue for speaker similarity systems.

4.2 Objective Accent Results

The listening test provides interesting results which are grounded in human perception. However, in part due to budget and in part due to the constraints of such a test, there are limits to what we can learn from the results. To further explore what is causing the low CosyVoice 2 accent SMOS scores, and to expand past a single speaker and 60 utterances, objective metrics and visualisations are required.

Three further speakers were randomly selected from the MEAD dataset: 1 male speaker (M012) and 2 female speakers (W018, W033). To reduce compute time and costs, only one reference sentence, 001, was used. Results for speaker M007 in this section also exclusively used the utterances produced using reference 001. The reduction was chosen based on the minimal effect that the reference sentence had on the entropy results shown previously in Section 3.3. As with that experiment, each TTS system was used to synthesise 100 utterances for each target text/emotion combination. An embed-

ding was then extracted for each utterance using GenAID [Zhong et al., 2025]. UMAP [McInnes et al., 2020] was then applied to reduce the dimensions of these embeddings. This same process was applied to both the ground truth utterances from the MEAD dataset as well as 19 English speakers with distinct accents from the CommonAccent [Zuluaga-Gomez et al., 2023] and VCTK[2] databases. For CommonAccent, 50 samples were collected for each accent. Using multiple speakers was found to create unexpected groupings in the UMAP space due to GenAID capturing speaker information beyond just accent. Thus, a single speaker with 5 or more utterances was randomly selected. Speakers with poor recording quality were excluded. For VCTK, a speaker was randomly selected for each desired accent. The first 50 utterances were used for each speaker. This constitutes 25 unique sentences, each recorded with two different microphones.

This process allows the accent of each utterance to be visualised in a 2D space composed of the first 2 UMAP components. This space can be seen in Figures 4.2 and 4.3 which allows for visualisation of how target text impacts accent realisation. The results seem to support those from the listening test in Table 4.1 as well as the entropy results in Table 3.2 which showed different target texts can lead to changes in accent. Utterances which received high entropy but low accent SMOS, such as *CV2-021-Happy* or *CV2-026-Angry*, are spread further with clear patches of outliers. Beyond individual target texts, we see clear patterns at higher levels. All 3 CosyVoice 2 plots appear much more spread than their MaskGCT counterparts. The CosyVoice 2 *happy* points show particular uncertainty. This aligns well with both the listening test, where *happy* utterances received the lowest average accent SMOS, and with the entropy results in which *happy* utterances were the most uncertain. Comparison between previous results and Figures 4.2 and 4.3, shows that in the case of a mismatch, such as between the accent SMOS and average entropy of *CV2-021-Neutral*, the visualisation seems to better align with the entropy results. This is perhaps unsurprising given the source of both stems from GenAID, however it is important to be aware of such mismatches. Objective results should be grounded in human perception scores where possible.

The results in these figures support the idea that target text impacts accent hallucination. However, some phenomena, such as the mismatch in *021* scores or the apparent spread shown in the MaskGCT happy and neutral plots, which doesn't show up

[2]https://datashare.ed.ac.uk/handle/10283/3443

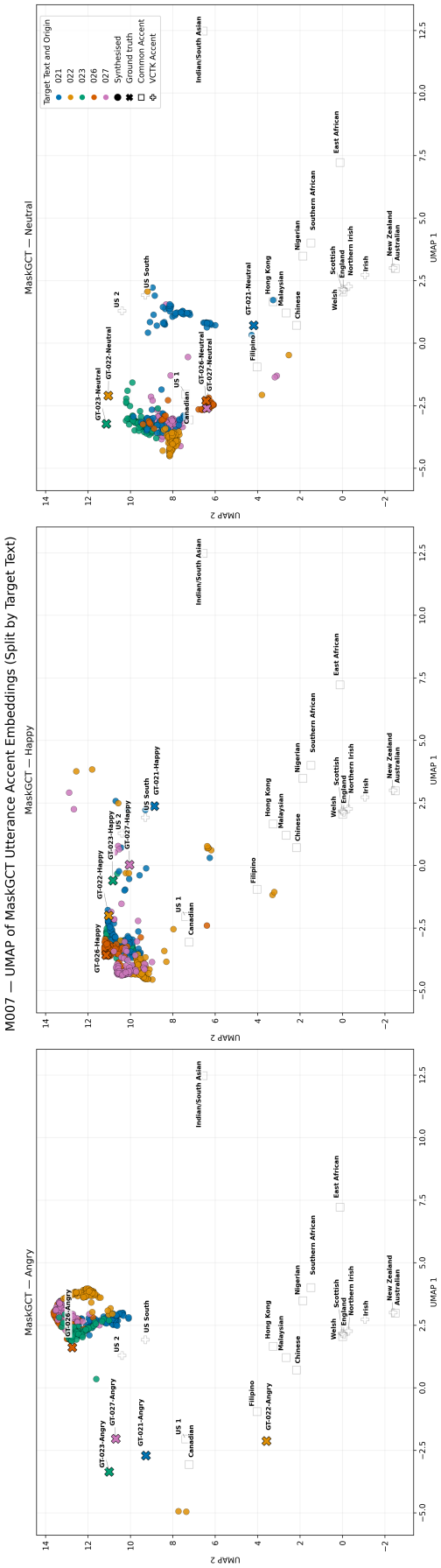


Figure 4.2: UMAP plots depicting MaskGCT utterances synthesised using reference speech from speaker M007. Each point represents the accent embedding of an utterance derived using GenAID [Zhong et al., 2025] and reduced to 2 dimensions. Utterances are split by the target text used to generate them. Ground truth accent points are included generated using accents from CommonAccent and VCTK.

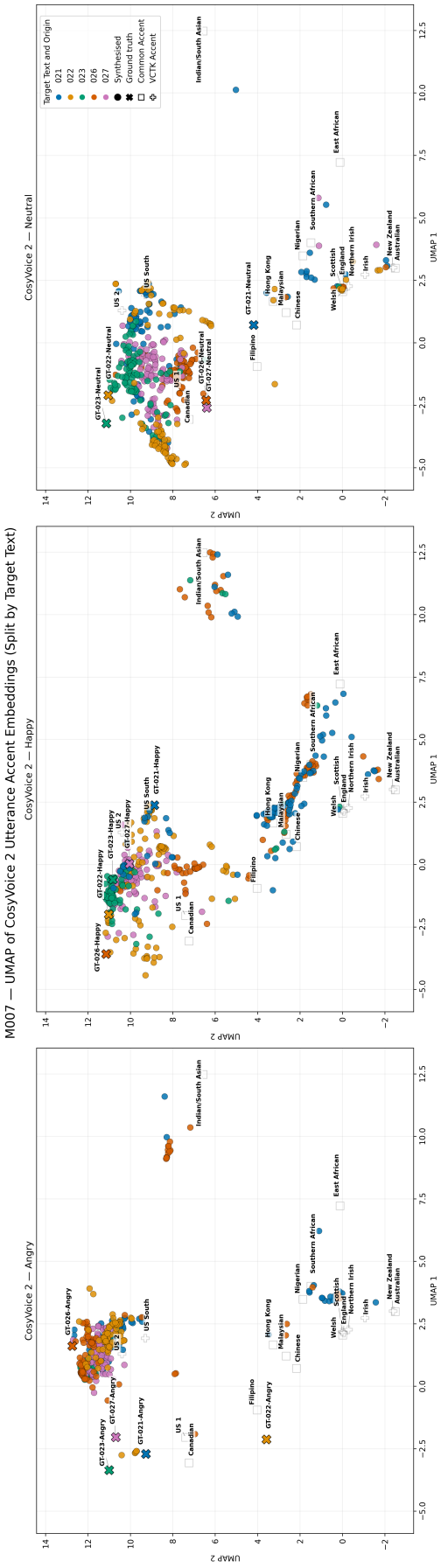


Figure 4.3: UMAP plot depicting CosyVoice 2 utterances synthesised using reference speech from speaker M007. Each point represents the accent embedding of an utterance derived using GenAID [Zhong et al., 2025] and reduced to 2 dimensions. Utterances are split by the target text used to generate them. Ground truth accent points are included generated using accents from CommonAccent and VCTK.

in other results, can be explored further by adding in the ground truth accent points taken from CommonAccent and VCTK. Figures 4.4-4.7 add these labels, as well as the plots for the 3 additional MEAD speakers. These immediately help to explain the phenomena mentioned previously. The spread seen in the MaskGCT M007 happy and neutral plots occurs almost entirely within a cluster of various North American accents. With this context, there seems to be no mismatch between the entropy, SMOS and visualisation. As predicted, MaskGCT does not seem to hallucinate accents for any emotion or target text. Figure 4.4 also labels a selection of the utterances included in the listening test. Using these labels, the mismatch between SMOS and objective measures for *CV2-021-Neutral* can be partially explained. The specific utterance used in the listening test belongs to a large cluster of neutral utterances among the North American accents. Listening to *GT-021-Neutral*, it does not seem distinct in accent from the other M007 ground truth utterances. Its position in Figures 4.3-4.2 seems to be an error caused either by GenAID or the UMAP mapping. This in turn led to the CosyVoice 2 utterance having a low cosine similarity and being selected for the listening test, despite it being typical in reality. This error reaffirms the need for multiple forms of evaluation, including subjective testing. Using only accent cosine similarity, it may have seemed that *CV2-021-Neutral* utterances show more accent hallucination than any other condition, which is not the case.

Further evidence for agreement between Figure 4.4 and accent SMOS can be found in the other labelled listening test utterances. For MaskGCT, utterance *022* is given for all emotions. All 3 lie within the North American accent cluster. This matches the high SMOS score each utterance received. The slightly lower SMOS score for *MGCT-022-Angry* is interesting as it aligns with the increased distance between the synthesised utterance and the ground truth, which itself is an outlier. The combination of these results suggests that speaker M007's accent may indeed differ slightly for this ground truth utterance. 4 listening test utterances are labelled in the M007 CosyVoice 2 plot. Aside from *CV2-021-Neutral*, the remaining utterances (*CV2-021-Happy/Angry* and *CV2-022-Neutral*) show clear deviations from their respective clusters. This again aligns with their accent SMOS scores which were low for all 3 (1.17, 1.33, 1.67). Each also belongs to its own mini-cluster. Listening to these utterances confirms their position in the UMAP space is appropriate. *CV2-021-Happy* and *CV2-021-Angry* seem to have African accents of some type, while *CV2-022-Neutral* has a clear British accent. Notably, these mini-clusters are largely populated by utterances of the same emotion.

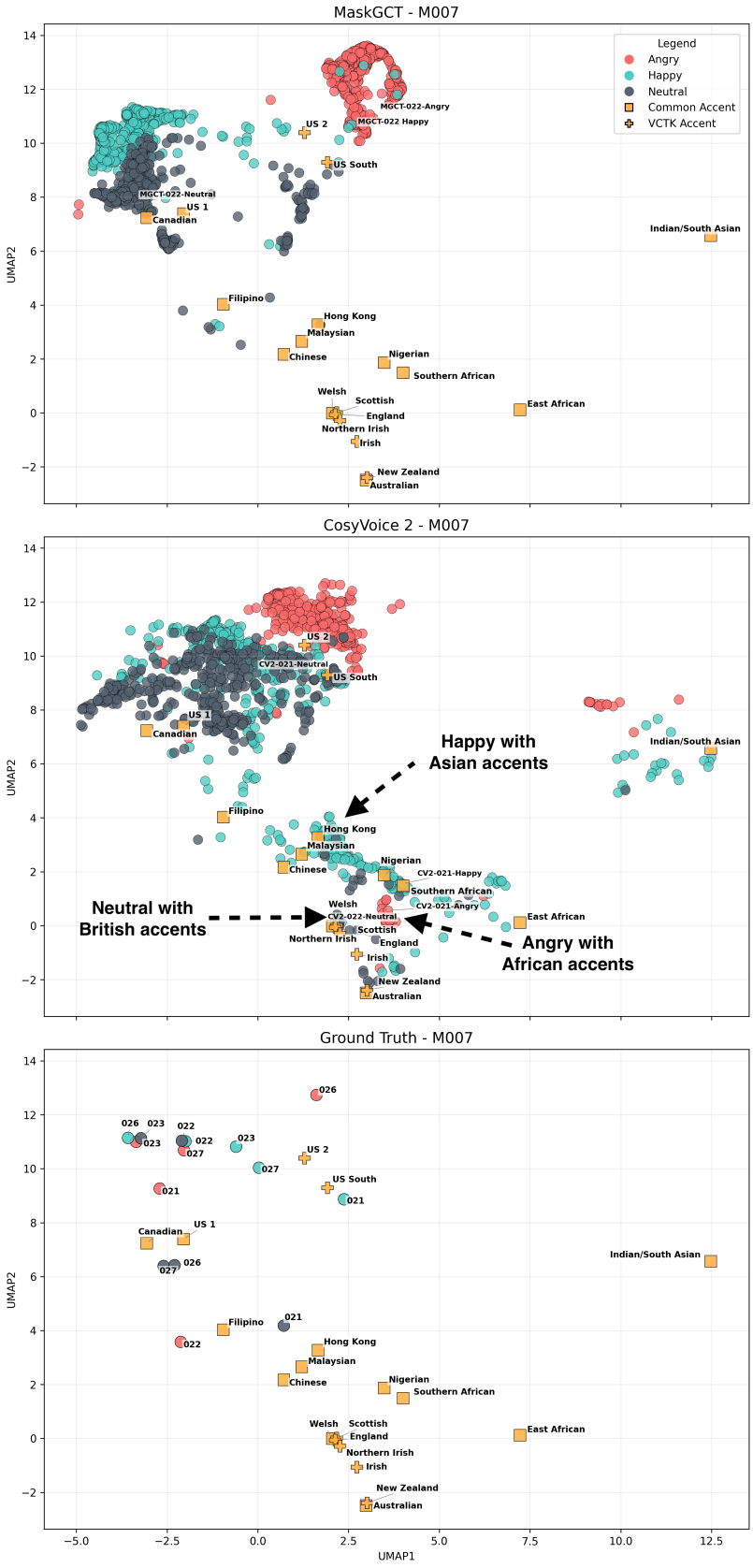


Figure 4.4: UMAP plots for speaker M007 showing how emotional utterance accents cluster. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth accent points are included from the CommonAccent and VCTK databases.

Figure 4.5: UMAP plots for speaker M012 showing how emotional utterance accents cluster. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth accent points are included from the CommonAccent and VCTK databases.

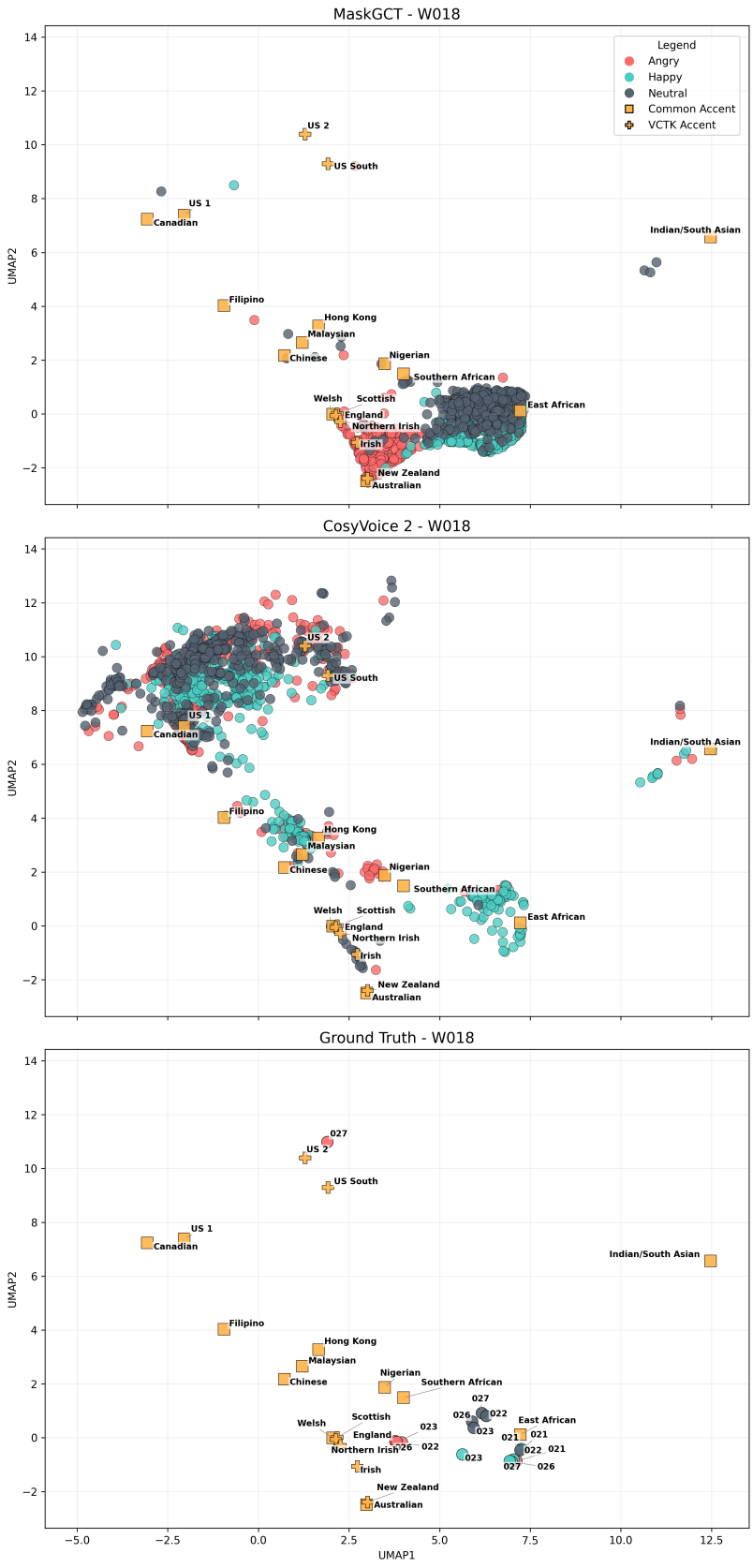


Figure 4.6: UMAP plots for speaker W018 showing how emotional utterance accents cluster. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth accent points are included from the CommonAccent and VCTK databases.

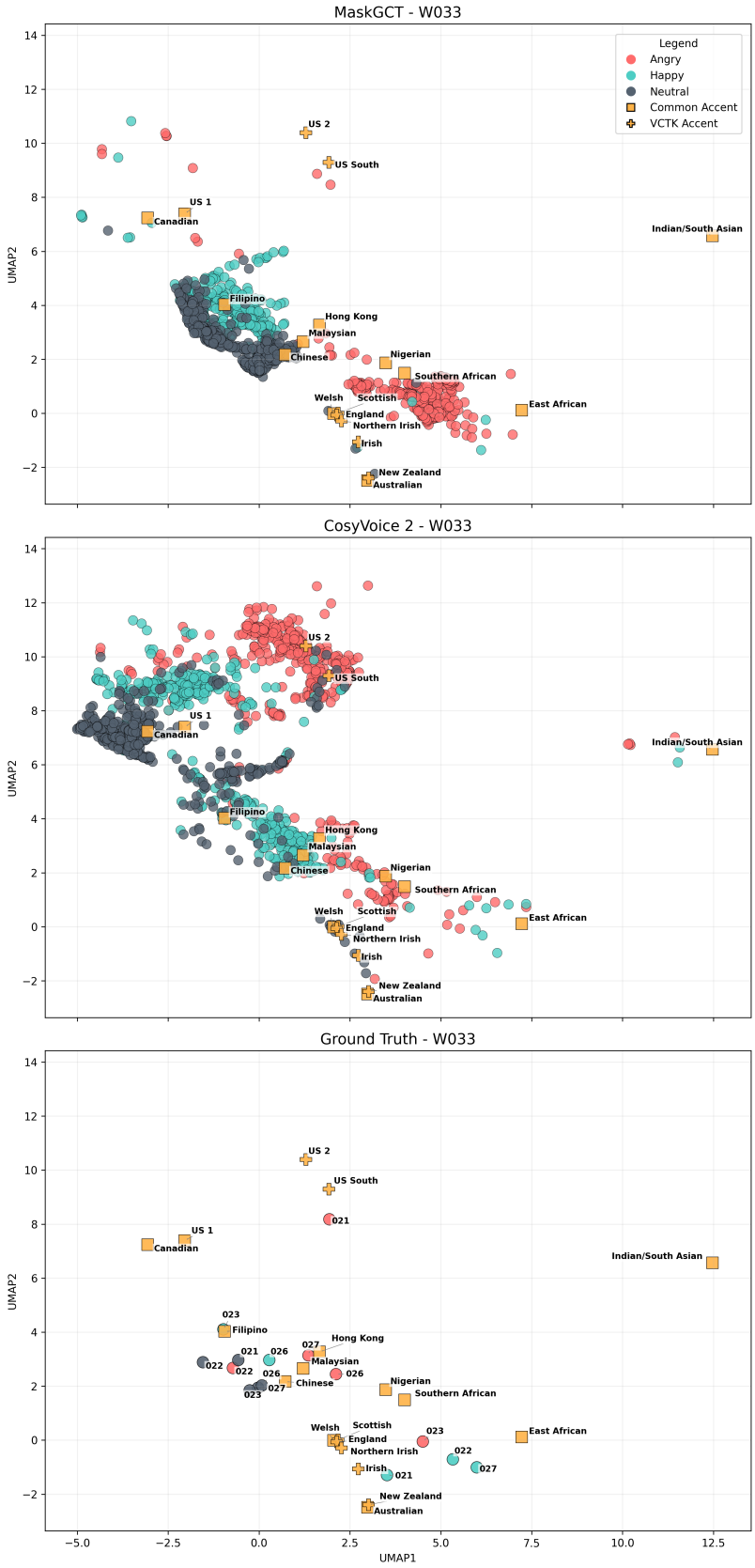


Figure 4.7: UMAP plots for speaker W033 showing how emotional utterance accents cluster. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth accent points are included from the CommonAccent and VCTK databases.

Speaker	Emotion	Cosine Distance to Centroid by System		
		Ground Truth	CosyVoice 2	MaskGCT
M007	Angry	0.017	**0.032**	0.017
	Happy	0.017	**0.055**	0.010
	Neutral	0.034	**0.035**	0.026
M012	Angry	0.019	**0.038**	0.014
	Happy	0.014	**0.058**	0.017
	Neutral	0.016	**0.033**	0.013
W018	Angry	0.033	**0.039**	0.037
	Happy	0.014	**0.052**	0.020
	Neutral	0.017	**0.029**	0.024
W033	Angry	0.020	**0.039**	0.023
	Happy	0.028	**0.054**	0.029
	Neutral	0.019	**0.042**	0.033

Table 4.3: Average cosine distance between an utterance and its centroid. Bold represents system with highest distance.

It seems that rather than accent hallucination occurring exclusively in utterances with emotional input, it instead can occur for instructions of any emotion. The input content does however guide how accent hallucination is realised.

Expanding beyond speaker M007, similar patterns can be seen for CosyVoice 2 utterances across speakers M012, W018 and W033. While the reference speaker does have a slight effect, the impact of instruction content and reference utterance emotion is greater. CosyVoice 2 utterances appear much more spread across all conditions than their MaskGCT or ground truth counterparts. This spread is quantified in Table 4.3 which gives the average cosine distance between utterance embeddings and the centroid of their respective emotion. This is calculated using the original GenAID embedding, not its reduced UMAP form.

These results confirm that CosyVoice 2 utterances are spread further than their MaskGCT or ground truth equivalents with cosine distance being greater for every speaker and emotion. As can be seen in Figure **??**, CosyVoice 2 happy utterances are on average

spread further from their centroid than neutral or angry utterances. Meanwhile, angry and neutral utterances are very close in average cosine distance, suggesting that neutral utterance see accent hallucination at a similar rate to angry ones. This pattern occurs regardless of reference speaker accent. Figures 4.8-4.11 show this clearly by overlaying the relevant ground truth utterances onto the synthesised utterances for each speaker. MaskGCT utterances tend to cluster around their ground truth equivalents. In contrast, CosyVoice 2 utterances ignore the accent of the original speaker. As mentioned in Section 3.2, when using the instruction mode of CosyVoice 2, accent seems to default to GenAm regardless of emotion or speaker. For example, ground truth utterances for speaker M012 have an Indian/South Asian accent. This is completely ignored by CosyVoice 2. The synthesised utterances which imitate the true accent follow a similar pattern to the hallucinated Indian/South Asian accents in other speakers and thus could themselves be considered hallucinated.

While they do generally follow a pattern for CosyVoice 2, there is some distinction between speakers. Speaker W033 in particular appears to have fewer utterances clustered in the North American accent section, and more towards Filipino/Asian accents. There is evidence this may be caused by GenAID. Ground truth utterances are split between Asian and African languages, suggesting the accent recognition system had difficulties pinning the speaker down. However, regardless of the cause, this is an interesting result as it points to the reference speaker's accent in some way influencing or being present in the accent produced by CosyVoice 2, even if the speaker's accent is rarely replicated. The failure of the cosine distances to capture this suggest that, while it is a useful metric, scores should be validated using further measures such as visualisation and subjective evaluation.

Finally, there is some evidence that the emotion of an utterance impacts the accent recognition performance of GenAID. In both MaskGCT and the ground truth we see some clustering of utterances with the same emotion. Even if these clusters are still where expected in the accent space, it suggests that further disentanglement of accent from emotion may be required.

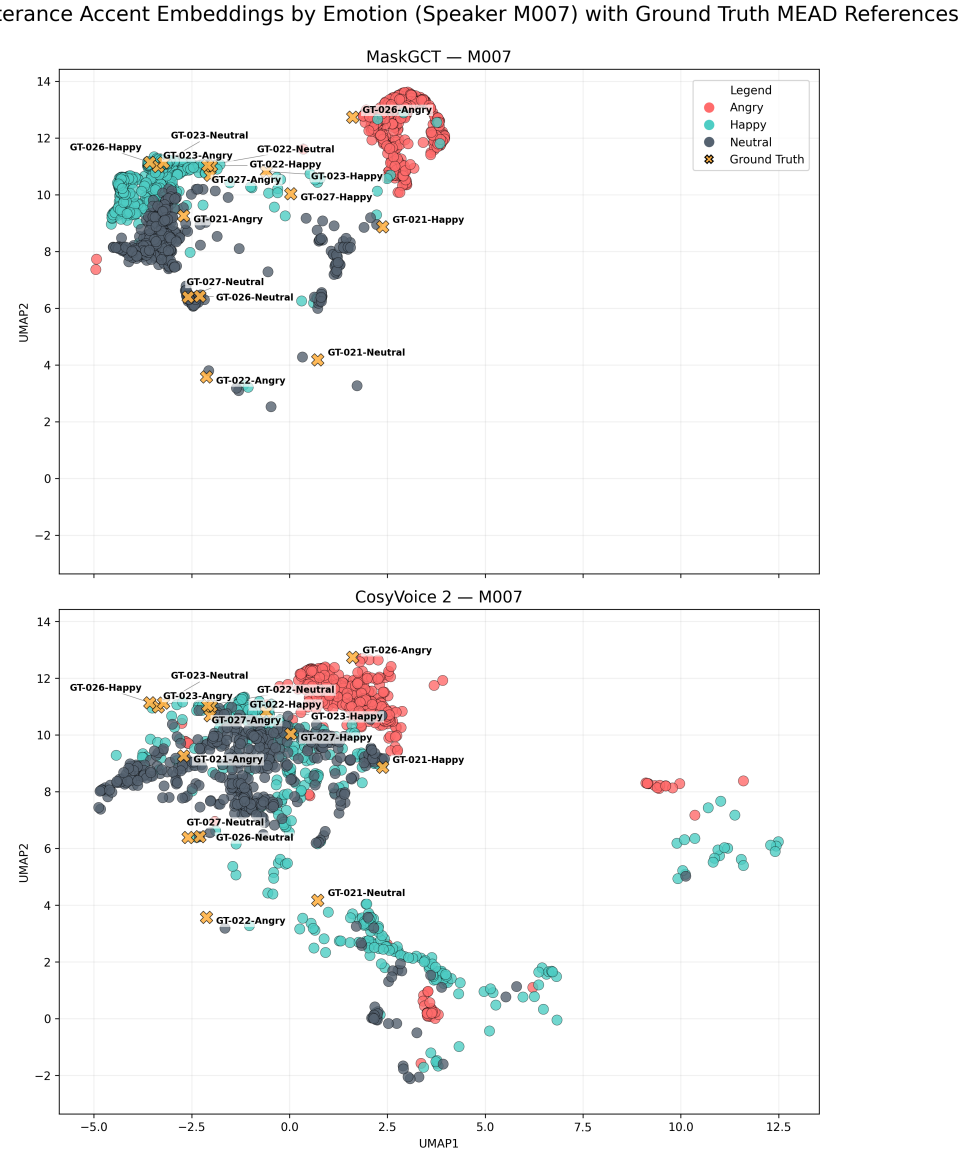


Figure 4.8: UMAP plots for speaker M007 showing how emotional utterance accents cluster for CosyVoice 2 and MaskGCT. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

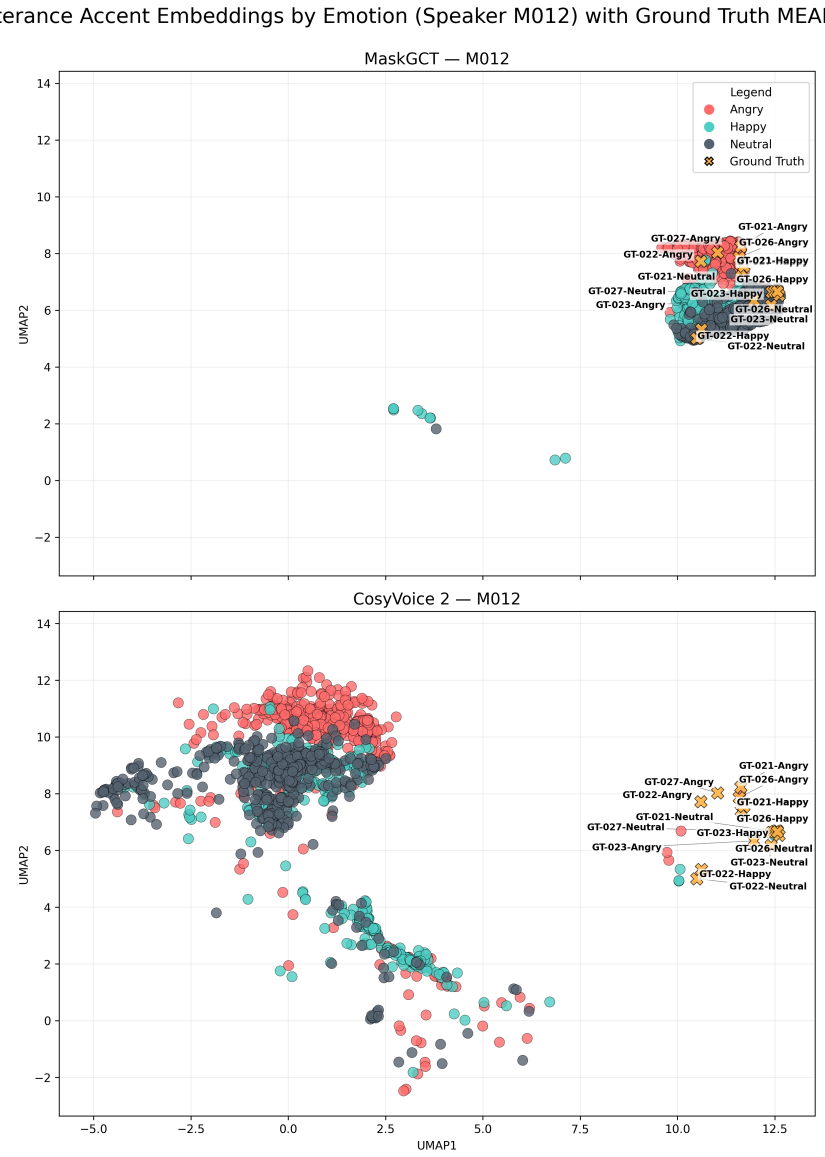


Figure 4.9: UMAP plots for speaker M012 showing how emotional utterance accents cluster for CosyVoice 2 and MaskGCT. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

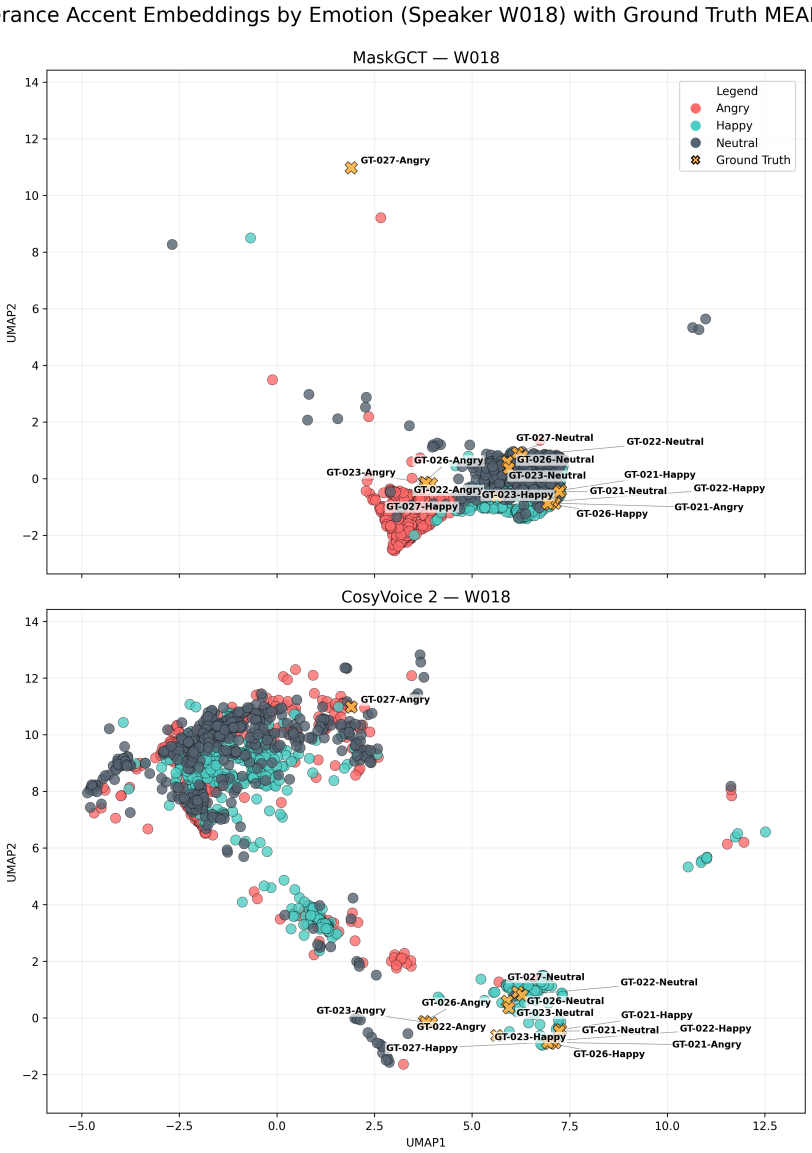


Figure 4.10: UMAP plots for speaker W018 showing how emotional utterance accents cluster for CosyVoice 2 and MaskGCT. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

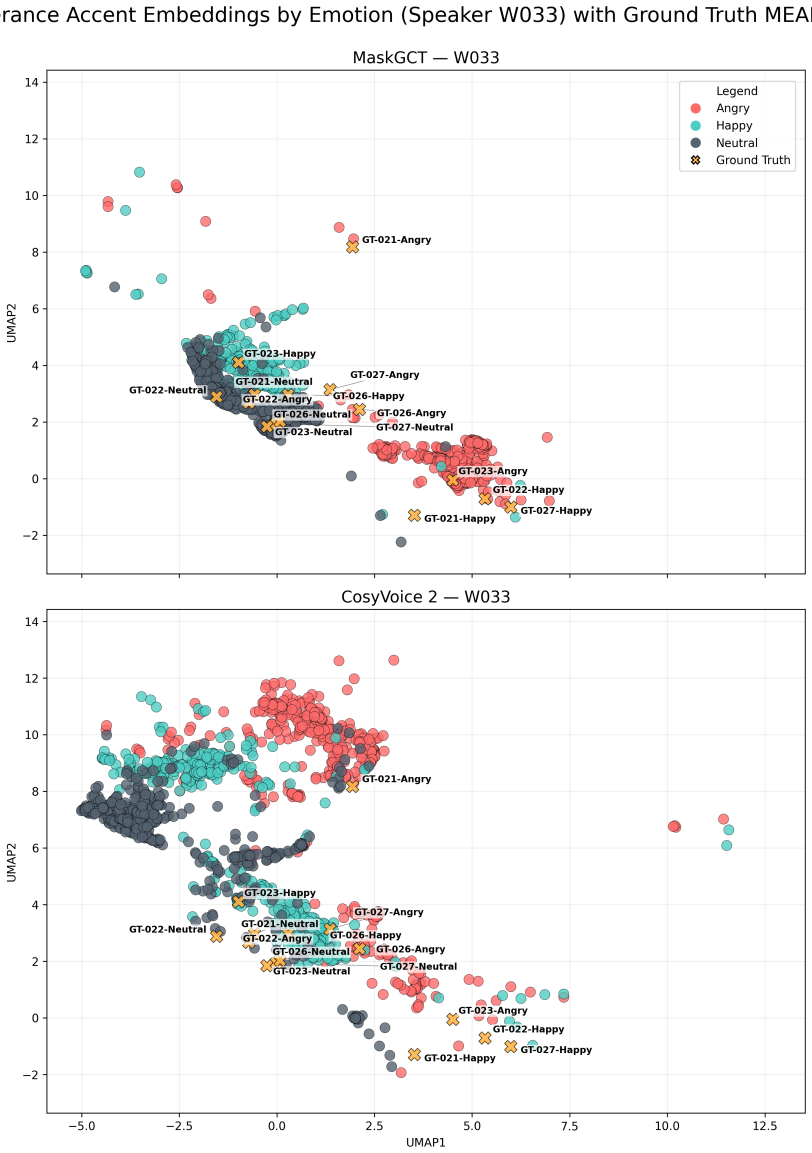


Figure 4.11: UMAP plots for speaker W033 showing how emotional utterance accents cluster for CosyVoice 2 and MaskGCT. Points represent utterance accent embeddings derived from GenAID [Zhong et al., 2025], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

4.3 Results Summary

Overall, it seems the results of the listening test in Section 4.1 may generalise to other speakers and utterances. CosyVoice 2 hallucinates accents across all 3 emotions. However, the extent to which it does so, and which accents are hallucinated, is controlled in part by the emotion of the reference utterance and instruction. Results from the listening test, cosine distance, visualisation and average entropy all show that *happy* emotional input leads to hallucination most frequently. The original hypothesis, that only *non-neutral* utterances will see accent hallucination, is thus falsified. Similarly, the hypothesis that CosyVoice 2 *angry* and *happy* utterances will underperform MaskGCT but performance on *neutral* utterances will not differ significantly, while technically not falsified by the listening test, does not seem correct. *Neutral* input results in accent hallucination at a similar rate to *angry* input. The nature of this hallucination does differ between the emotions.

4.4 Emotion Results

The third hypothesis in Section 3.6 posited that there would be no significant difference between the performance of MaskGCT and CosyVoice 2 in emotion SMOS. Just as with accent SMOS, this was measured in the listening test using a 1-5 scale with participants asked to rate how similar a synthesised utterance was to the ground truth equivalent in terms of emotion. The questions were part of the same listening test as the accent SMOS and used the same utterances. Table 4.4 presents the results of this test. CosyVoice 2 outperforms MaskGCT on most *angry* and *neutral* utterances. However, performance is much closer for *happy* utterances, with MaskGCT bettering CosyVoice 2. This can be seen more clearly in Table 4.5 which gives the scores averaged across each target text as well as the p-values derived from an unpaired t-test. We see here that there is a significant difference between CosyVoice 2 and MaskGCT for neutral utterances but not for *angry* or *happy* utterances, although the *angry* results are very close to significant. These results falsify the hypothesis. There is a clear difference between CosyVoice 2 and MaskGCT in terms of emotion SMOS for both *angry* and neutral utterances.

It is noteworthy that the results here mimic the accent SMOS scores for CosyVoice 2, with *angry* and *neutral* utterances outperforming *happy* ones overall. For CosyVoice

System	Emotion	Emotion SMOS by Target Text				
		021	**022**	**023**	**026**	**027**
CosyVoice 2	**Angry**	3.3 ± 1.4	**3.7 ± 1.02**	3.52 ± 1.12	**3.83 ± 1.15**	3.48 ± 1.56
	Happy	*1.96 ± 1.22*	*2.83 ± 1.03*	**4.04 ± 1.02**	*3.13 ± 1.32*	**3.65 ± 1.19**
	Neutral	**3.65 ± 1.07**	3.3 ± 1.29	*3.39 ± 1.34*	3.17 ± 1.56	*3.13 ± 1.32*
MaskGCT	**Angry**	*1.88 ± 1.31*	*1.67 ± 1.02*	**3.96 ± 1.07**	2.25 ± 1.46	3.13 ± 1.17
	Happy	**3.5 ± 1.31**	**3.83 ± 1.06**	3.79 ± 1.11	**3.25 ± 1.11**	*2.83 ± 1.35*
	Neutral	2.33 ± 1.22	2.25 ± 1.24	*3.04 ± 1.36*	*1.63 ± 1.04*	**3.38 ± 1.34**

Table 4.4: Emotion SMOS and standard deviation for CosyVoice 2 and MaskGCT across *angry*, *happy* and *neutral* utterances. The listening test was carried out with 29 L1 English speakers. Bold represents a system's highest scoring emotion for that target text, italics represents the lowest.

System	Average SMOS by Emotion		
	Angry	**Happy**	**Neutral**
CosyVoice 2	3.52 ± 0.20	3.13 ± 0.80	3.30 ± 0.20
MaskGCT	2.58 ± 0.95	3.44 ± 0.41	2.53 ± 0.69
p-value	0.0623	0.4628	**0.0434**

Table 4.5: Emotion SMOS averaged across target texts with p-values for determining significance. Bold indicates a significant result.

2 however, there does not seem to be a correlation between the accent and emotion SMOS scores with a Pearson correlation of $r = 0.073$. MaskGCT does see a correlation. Its Pearson correlation of $r = 0.393$ indicates a weak positive correlation, as accent SMOS increases so does emotion SMOS. This may be due to how utterances were initially chosen based on accent cosine similarity. There doesn't seem to be much shift in accent for MaskGCT utterances so low accent cosine similarity may result in picking out utterances that are generally unnatural sounding, resulting in low scores for both accent and emotion SMOS.

To visualise the differences in emotion realisation. Utterances were put through emotion2vec [Ma et al., 2023] in order to produce emotion embeddings. These embeddings were then fit to a UMAP space, along with the ground truth utterances, allowing the first 2 UMAP components of each embedding to be plotted. This was carried out for each speaker using the same utterances as in the accent visualisations. The results of this process are presented in Figures 4.12-4.15. A selection of the utterances used in the listening test are labelled. Immediately we can see much tighter clustering for angry and neutral emotions in CosyVoice 2 than in MaskGCT. This aligns with the SMOS scores. Looking at the labelled listening test utterances, SMOS scores again seem in line with the visualisation. Low scoring utterances for *MGCT-022-Angry* and *MGCT-026-Neutral* both appear to be outliers, far from their respective clusters. This lends credence to the idea that, for MaskGCT, choosing utterances with low accent cosine similarity results in choosing utterances that are poorly synthesised in general. For CosyVoice 2, *CV2-021-Happy* and *CV2-022-Happy* are both lower scoring utterances which are some distance from their ground truth equivalents while *CV2-022-Neutral* scores higher and is placed right next to its ground truth equivalent.

These results closely match the SMOS scores just as the subjective and objective results were aligned when measuring accent similarity. However, if we were to focus only one aspect, i.e. only emotion or only accent, we would miss key information about system performance. Using only emotion metrics, CosyVoice 2 would seem to outperform MaskGCT in terms of speaker imitation. When viewed in conjunction with the accent results, it is obvious that this is not the case. This highlights the importance of avoiding catch-all metrics and evaluating as many aspects of speech and speaker similarity as possible in order to to accurately capture model performance.

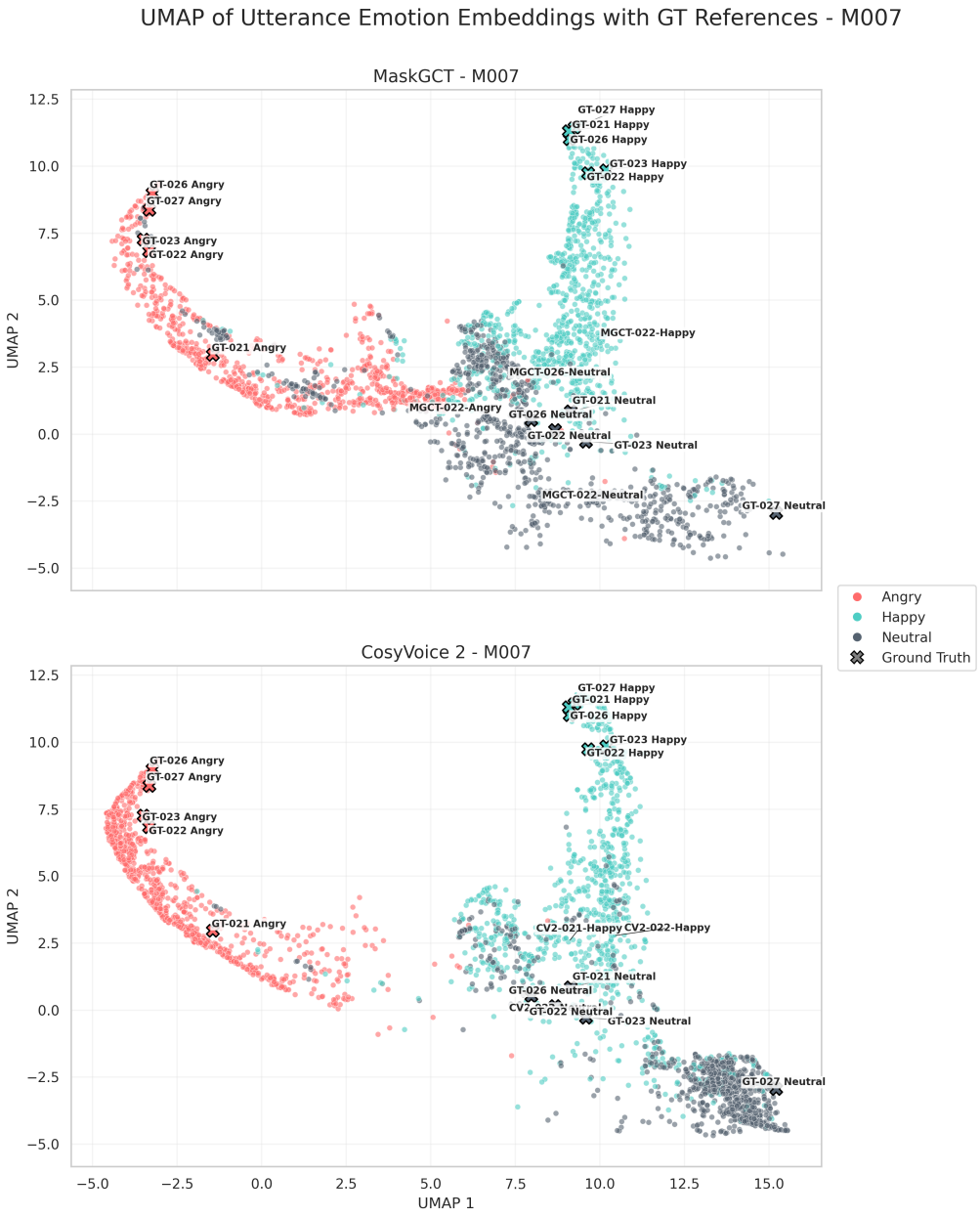


Figure 4.12: UMAP plots for speaker M007 showing how emotional utterances cluster for CosyVoice 2 and MaskGCT. Points represent utterance emotion embeddings derived from emotion2vec [Ma et al., 2023], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

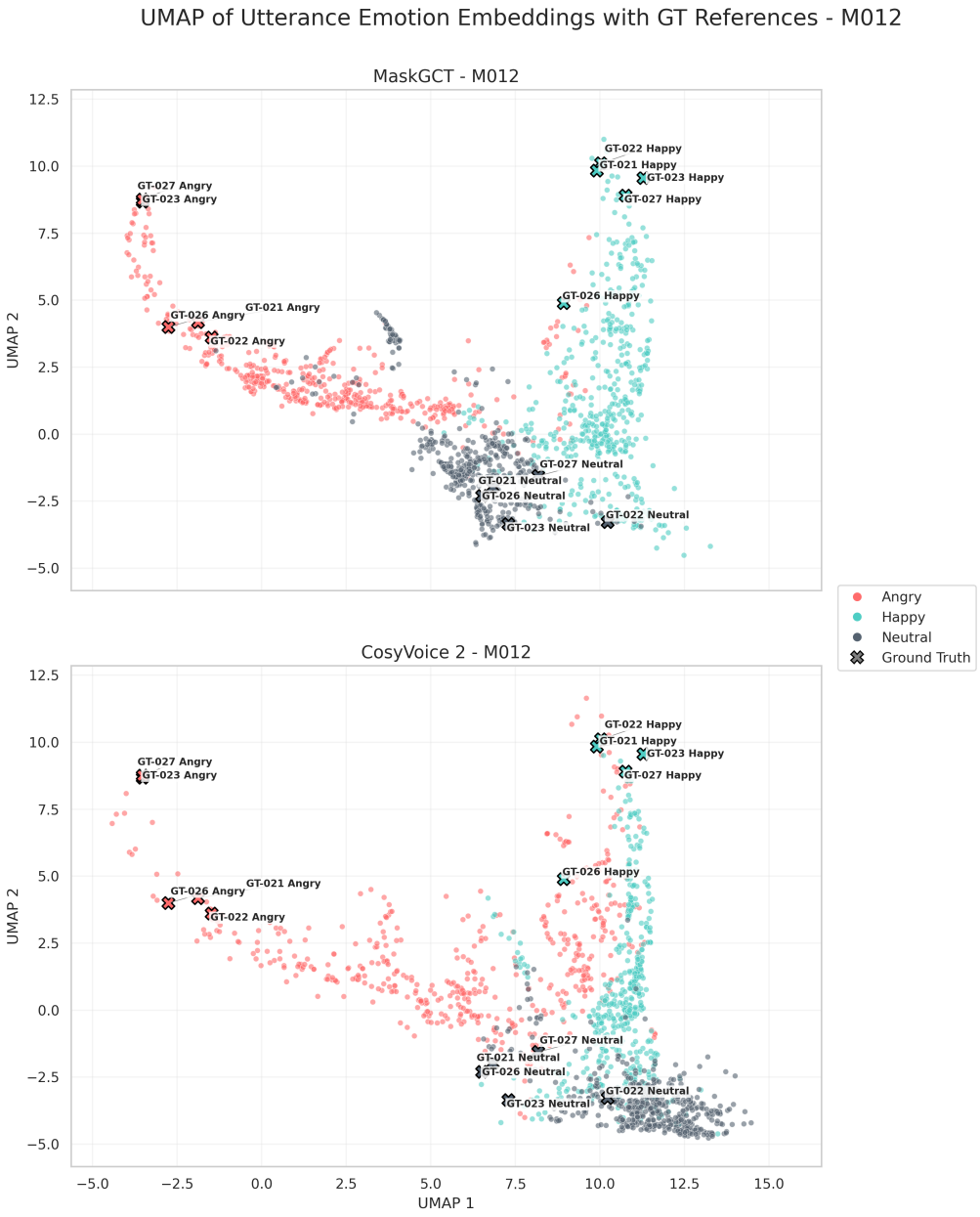


Figure 4.13: UMAP plots for speaker M012 showing how emotional utterances cluster for CosyVoice 2 and MaskGCT. Points represent utterance emotion embeddings derived from emotion2vec [Ma et al., 2023], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

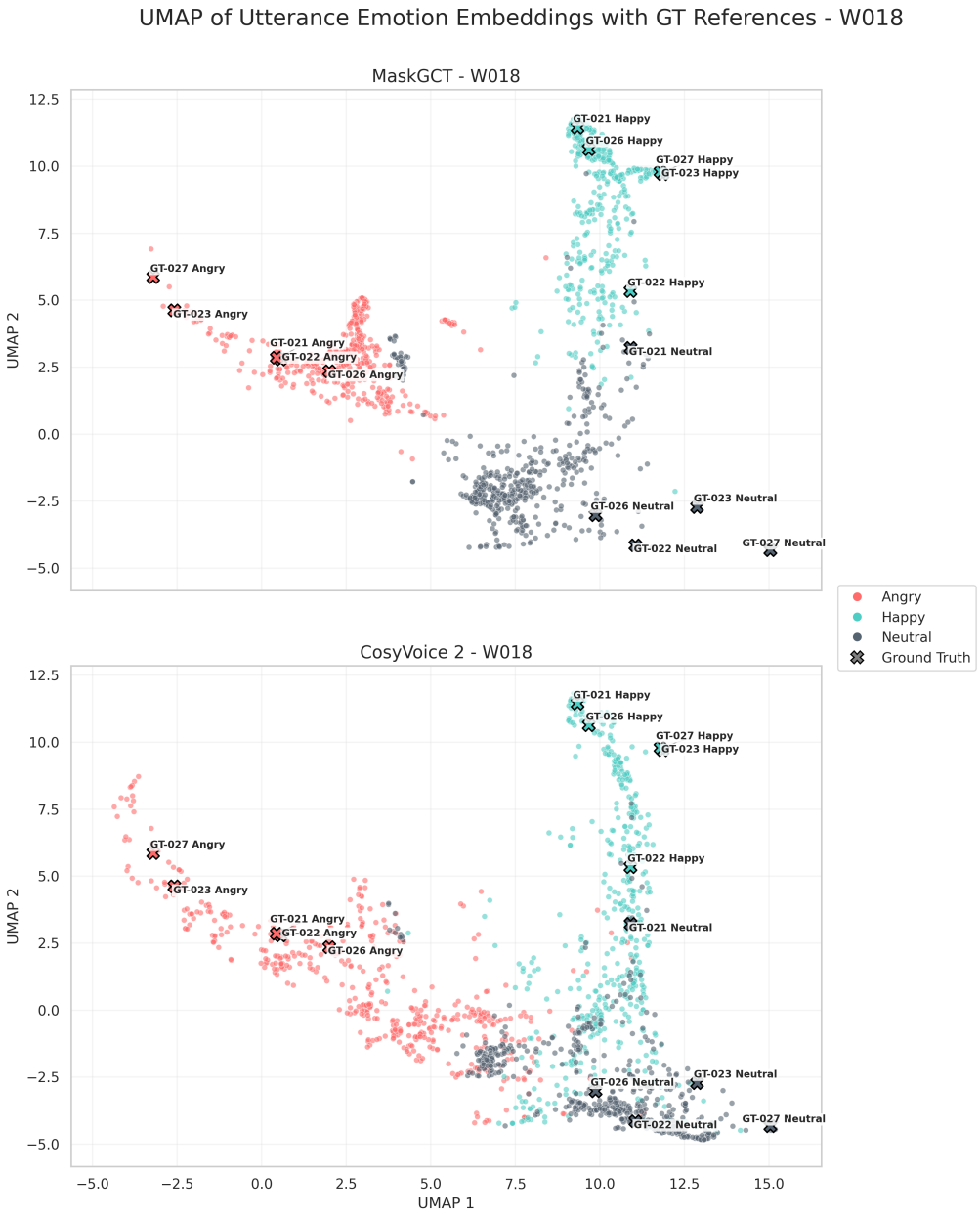


Figure 4.14: UMAP plots for speaker W018 showing how emotional utterances cluster for CosyVoice 2 and MaskGCT. Points represent utterance emotion embeddings derived from emotion2vec [Ma et al., 2023], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

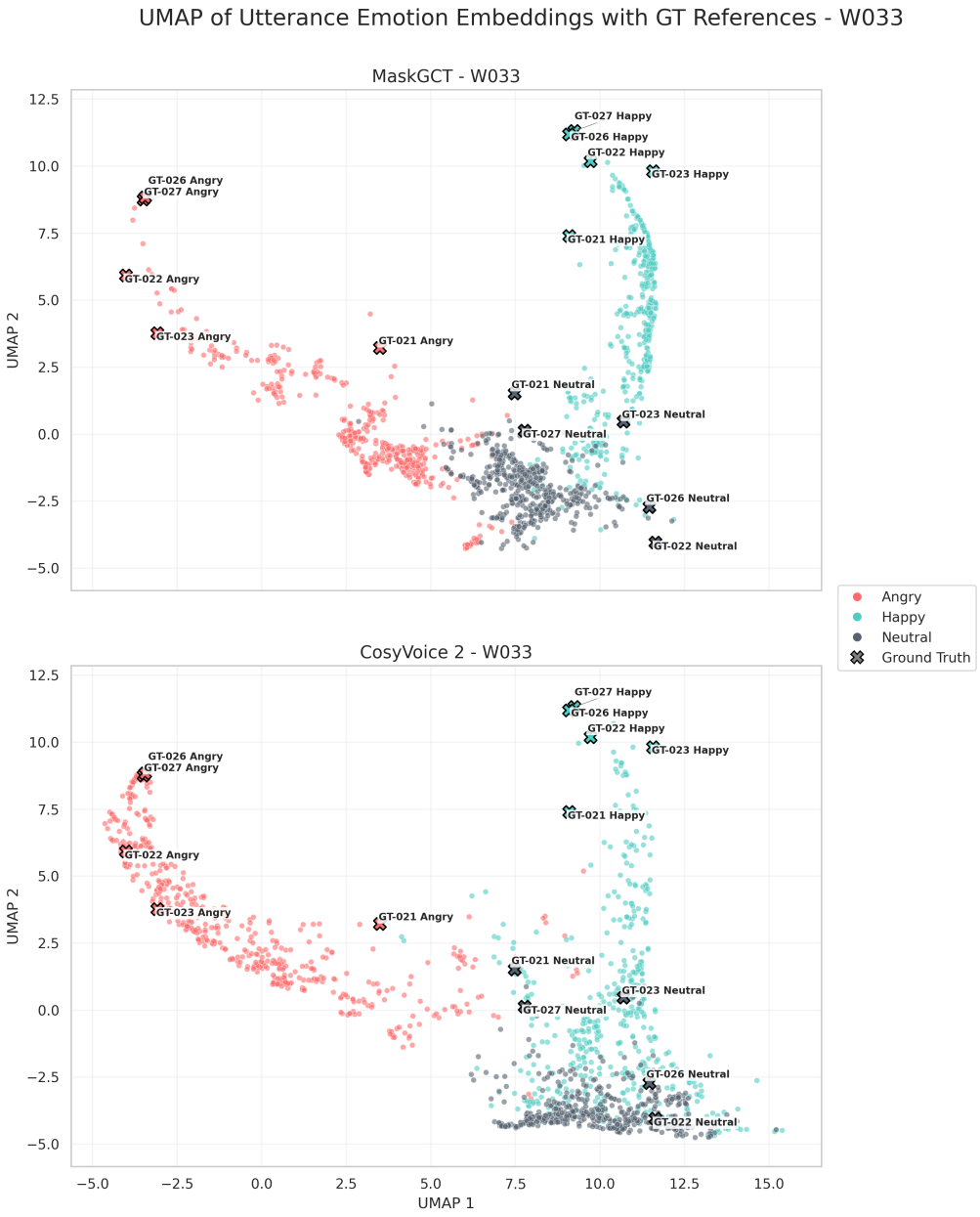


Figure 4.15: UMAP plots for speaker W033 showing how emotional utterances cluster for CosyVoice 2 and MaskGCT. Points represent utterance emotion embeddings derived from emotion2vec [Ma et al., 2023], reduced to 2 dimensions. Ground truth points are included from the original speaker in the MEAD database.

Chapter 5

General Discussion

Naturally with such a novel phenomenon, research has prompted more questions than answers. It is therefore important to take stock and focus on the key findings from the results. Discussion in this chapter turns to the overarching questions posed in this paper and the answers provided by analysis of both the subjective and objective results.

5.1 What is the Nature of Accent-Emotion Entanglement in CosyVoice 2?

The results presented in Sections 4.1-4.2 demonstrate that the link between accent hallucination and emotional input in CosyVoice 2 differs slightly from the initial hypothesis. Accent hallucination occurs in all three emotion conditions test: *angry, happy* and *neutral*. However, each emotion leads to different patterns in the accents that are hallucinated. *Happy* emotional input lead to the greatest spread with large clusters receiving hallucinated African, Filipino, Chinese or Indian/South Asian accents. African accents were also commonly hallucinated for utterances with *angry* emotional input. Meanwhile, *neutral* emotional input led to a cluster of utterances hallucinating British accents. These trends were broadly the same for each speaker, however the fact that minimal differences that were present suggests that the accent of the original speaker is in some way still leaking through to the synthesised utterances. The target text of each utterance was also shown to impact realisation of the accent hallucination in CosyVoice 2. Different target texts led to differing patterns with some, having a much greater spread than others, such as *CV2-021* vs. *CV2-027*. Given the content of each target text was emotionally neutral, with no clear link to any accent, it is not clear why

this occurred. Accent-emotion entanglement was only seen in CosyVoice 2, ground truth and MaskGCT utterances received higher accent SMOS scores and were tightly clustered in the UMAP accent space.

5.2 How Should we Evaluate Similarity Between Speakers?

Comparison of the results in Sections 4.1 and 4.2 confirmed the benefit of using a diverse range of metrics and methods when measuring speaker similarity. A core belief of this paper was that any objective results should be validated using subjective results. Human-centric measures such as SMOS scores ground any findings in real world perceptual differences. This allows for quick and easy confirmation or rejection of objective results which seem out of place. For example, the specific utterance chosen to represent *CV2-021-Neutral* in the listening test had a very low cosine similarity (0.70) to its ground truth equivalent. This was intended to signify that the synthesised utterance was atypical and thus may show accent hallucination. However, the UMAP visualisation disputed this, showing *CV2-021-Neutral* was situated exactly where expected, clustered with other *neutral* utterances amongst the North American accents. Accent SMOS scores helped to solve this disagreement. *CV2-021-Neutral* scored quite highly (4.13), suggesting an issue with the cosine similarity, not the UMAP visualisation. Without the subjective results from the listening test, the true nature of *CV2-021-Neutral* may not have been confirmed.

Subjective evaluations allow us to validate that any metric is measuring or approximating what it's supposed to. Many TTS papers use speaker similarity metrics which are poorly defined or described. Although TTS systems results may be placed alongside ground truth or vocoder resynthesised results, it is often not obvious what speaker similarity captures and how any differences would actually be realised perceptually. Placing such metrics alongside subjective results, as is done in [Wang et al., 2024], gives a clear guide on how to interpret them and what they mean for speech quality. In Section 4.1, accent SMOS results aligned very well with the average entropies of utterance accent probability distributions, high SMOS scores often meant lower average entropy. Validating the entropy results in this way has the same effect as the results comparison in Wang et al (2024), allowing the reader to place the results of a novel

objective metric in the context of an easily understandable listening test. Similarly, using a wider range of objective measures and visualisations can help to paint a clearer picture of speech quality and confirm that certain results aren't outliers. The cosine similarity, UMAP visualisations and cosine distance from centroid used in this paper all aligned to clearly illustrate the presence of accent hallucination in CosyVoice 2 utterances. In conclusion, it is important to provide a diverse range of evaluation metrics, both subjective and objective in order to validate and clarify any results.

5.3 Why Does Accent Hallucination Occur in CosyVoice 2?

Without knowledge of the dataset used to train CosyVoice 2, it is difficult to answer this question. Considering the similarities between MaskGCT and CosyVoice 2, as well as the differing results, suggest that the issue does not stem from the architecture of the model. Instead, the root cause may lie in the instruction fine-tuning data. Although we know little about the data, we know it is labelled with emotion automatically using SenseVoice [An et al., 2024]. It could be that SenseVoice is mistaking accent features for emotional ones, labelling, for example, neutral African-accented as angry because of the accent. This would lead to certain accents being over-represented in the instruction dataset and could lead to unexpected trends such as accent-emotion entanglement. Alternatively, the accent hallucination could stem from imbalance in the dataset. It might be that, as an example, British speakers are truly over-represented in the *neutral* instruction dataset. This would greatly increase the likelihood of CosyVoice 2 hallucinating a British accent for utterances with *neutral* emotional input without nay actual systems being at fault. In truth, as stated, it is very difficult to pin down the root cause without access to the training data for CosyVoice 2. This highlights one of the many problems with keeping such information private.

Chapter 6

Future Work and Conclusions

This paper introduced a novel concept: accent-emotion entanglement. This phenomenon was defined as accent hallucination caused or guided by the emotion content of the input provided to a TTS system. CosyVoice 2 and MaskGCT were tested in order to investigate the issue. It was found that no accent hallucination occurred in utterances produced by MaskGCT. CosyVoice 2, on the other hand, frequently hallucinated accents. Both subjective and objective metrics demonstrated that this hallucination was indeed guided by the emotion of the input provided. Utterances provided with angry, happy or neutral input all produced differing patterns of accents. CosyVoice 2 thus suffers from accent-emotion entanglement.

Future work will further investigate the nature of this entanglement. One avenue of research could focus on differing the emotion of the reference speech and emotion. This could give a better understanding of exactly how each element of the input affects accent hallucination. To help solve the problem, CosyVoice 2 could also be fine-tuned using a known dataset well-balanced for accent and emotion. The results of this fine-tuning could help point to the overall cause of the accent-emotion entanglement.

Bibliography

[An et al., 2024] An, K., Chen, Q., Deng, C., Du, Z., Gao, C., Gao, Z., Gu, Y., He, T., Hu, H., Hu, K., Ji, S., Li, Y., Li, Z., Lu, H., Luo, H., Lv, X., Ma, B., Ma, Z., Ni, C., Song, C., Shi, J., Shi, X., Wang, H., Wang, W., Wang, Y., Xiao, Z., Yan, Z., Yang, Y., Zhang, B., Zhang, Q., Zhang, S., Zhao, N., and Zheng, S. (2024). FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. arXiv:2407.04051 [cs].

[Anastassiou et al., 2024] Anastassiou, P., Chen, J., Chen, J., Chen, Y., Chen, Z., Chen, Z., Cong, J., Deng, L., Ding, C., Gao, L., Gong, M., Huang, P., Huang, Q., Huang, Z., Huo, Y., Jia, D., Li, C., Li, F., Li, H., Li, J., Li, X., Li, X., Liu, L., Liu, S., Liu, S., Liu, X., Liu, Y., Liu, Z., Lu, L., Pan, J., Wang, X., Wang, Y., Wang, Y., Wei, Z., Wu, J., Yao, C., Yang, Y., Yi, Y., Zhang, J., Zhang, Q., Zhang, S., Zhang, W., Zhang, Y., Zhao, Z., Zhong, D., and Zhuang, X. (2024). Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. arXiv:2406.02430 [eess].

[Badlani et al., 2023] Badlani, R., Valle, R., Shih, K. J., Santos, J. F., Gururani, S., and Catanzaro, B. (2023). RAD-MMM: Multilingual Multiaccented Multispeaker Text To Speech. In *INTERSPEECH 2023*, pages 626–630. ISCA.

[Casanova et al., 2024] Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. (2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv:2406.04904 [eess].

[Chen et al., 2023] Chen, Y., Zheng, S., Wang, H., Cheng, L., Chen, Q., and Qi, J. (2023). An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. arXiv:2305.12838 [eess].

[Du et al., 2024a] Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., Gao, Z., and Yan, Z. (2024a). CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. arXiv:2407.05407 [cs].

[Du et al., 2024b] Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., Yu, F., Liu, H., Sheng, Z., Gu, Y., Deng, C., Wang, W., Zhang, S., Yan, Z., and Zhou, J. (2024b). CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. arXiv:2412.10117 [cs].

[He et al., 2025] He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., Wang, Y., Chen, K., Zhang, P., and Wu, Z. (2025). Emilia Extended. arXiv:2501.15907 [cs].

[Jiang et al., 2024] Jiang, Z., Liu, J., Ren, Y., He, J., Ye, Z., Ji, S., Yang, Q., Zhang, C., Wei, P., Wang, C., Yin, X., Ma, Z., and Zhao, Z. (2024). Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis. arXiv:2307.07218 [eess].

[Jiang et al., 2023] Jiang, Z., Ren, Y., Ye, Z., Liu, J., Zhang, C., Yang, Q., Ji, S., Huang, R., Wang, C., Yin, X., Ma, Z., and Zhao, Z. (2023). Mega-TTS: Zero-Shot Text-to-Speech at Scale with Intrinsic Inductive Bias. arXiv:2306.03509 [eess].

[Lyth and King, 2024] Lyth, D. and King, S. (2024). Natural language guidance of high-fidelity text-to-speech with synthetic annotations. arXiv:2402.01912 [cs].

[Ma et al., 2023] Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. (2023). emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. arXiv:2312.15185 [cs].

[McInnes et al., 2020] McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat].

[Walley-Jean, 2009] Walley-Jean, J. C. (2009). Debunking the Myth of the ”Angry Black Woman”: An Exploration of Anger in Young African American Women. *Black Women, Gender & Families*, 3(2):68–86.

[Wang et al., 2020] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020). MEAD: A Large-Scale Audio-Visual Dataset

for Emotional Talking-Face Generation. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12366, pages 700–717. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

[Wang et al., 2024] Wang, Y., Zhan, H., Liu, L., Zeng, R., Guo, H., Zheng, J., Zhang, Q., Zhang, X., Zhang, S., and Wu, Z. (2024). MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer. arXiv:2409.00750 [cs].

[Xin et al., 2021] Xin, D., Komatsu, T., Takamichi, S., and Saruwatari, H. (2021). Disentangled Speaker and Language Representations Using Mutual Information Minimization and Domain Adaptation for Cross-Lingual TTS. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6608–6612. ISSN: 2379-190X.

[Zhao et al., 2018] Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., and Gutierrez-Osuna, R. (2018). L2-ARCTIC: A Non-native English Speech Corpus. In *Interspeech 2018*, pages 2783–2787. ISCA.

[Zhong et al., 2025] Zhong, J., Richmond, K., Su, Z., and Sun, S. (2025). AccentBox: Towards High-Fidelity Zero-Shot Accent Generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

[Zhou et al., 2021] Zhou, K., Sisman, B., Liu, R., and Li, H. (2021). Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. ISSN: 2379-190X.

[Zhu et al., 2024] Zhu, X., Lei, Y., Li, T., Zhang, Y., Zhou, H., Lu, H., and Xie, L. (2024). METTS: Multilingual Emotional Text-to-Speech by Cross-Speaker and Cross-Lingual Emotion Transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1506–1518.

[Zuluaga-Gomez et al., 2023] Zuluaga-Gomez, J., Ahmed, S., Visockas, D., and Subakan, C. (2023). CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. arXiv:2305.18283 [cs].

Appendix A

Appendix A - MEAD Target Sentences

Below are the target sentences and corresponding utterance numbers used throughout the paper. They are taken from the MEAD dataset [Wang et al., 2020].

Common Sentences - Used as reference sentences throughout.

001. She had your dark suit in greasy wash water all year

002. Don't ask me to carry an oily rag like that

Generic Sentences - Used as target texts throughout.

021. Todd placed top priority on getting his bike fixed

022. One even gave my little dog a biscuit

023. I'll have a scoop of that exotic purple and turquoise sherbet

024. Land based radar would help with this task

026. His superiors had also preached this saying it was the way for eternal honor

027. It was not whatever tale was told by tails

028. No the man was not drunk he wondered how he got tied up with this stranger

029. No price is too high when true love is at stake

030. The revolution now under way in materials handling makes this much easier

Appendix B

Appendix B

The 13 accents present in GenAID [Zhong et al., 2025] probability distributions are as follows:

US

Canadian

Australian

South Asian

English

South African

Irish

Scottish

Filipino

Singaporean

Hong Kong

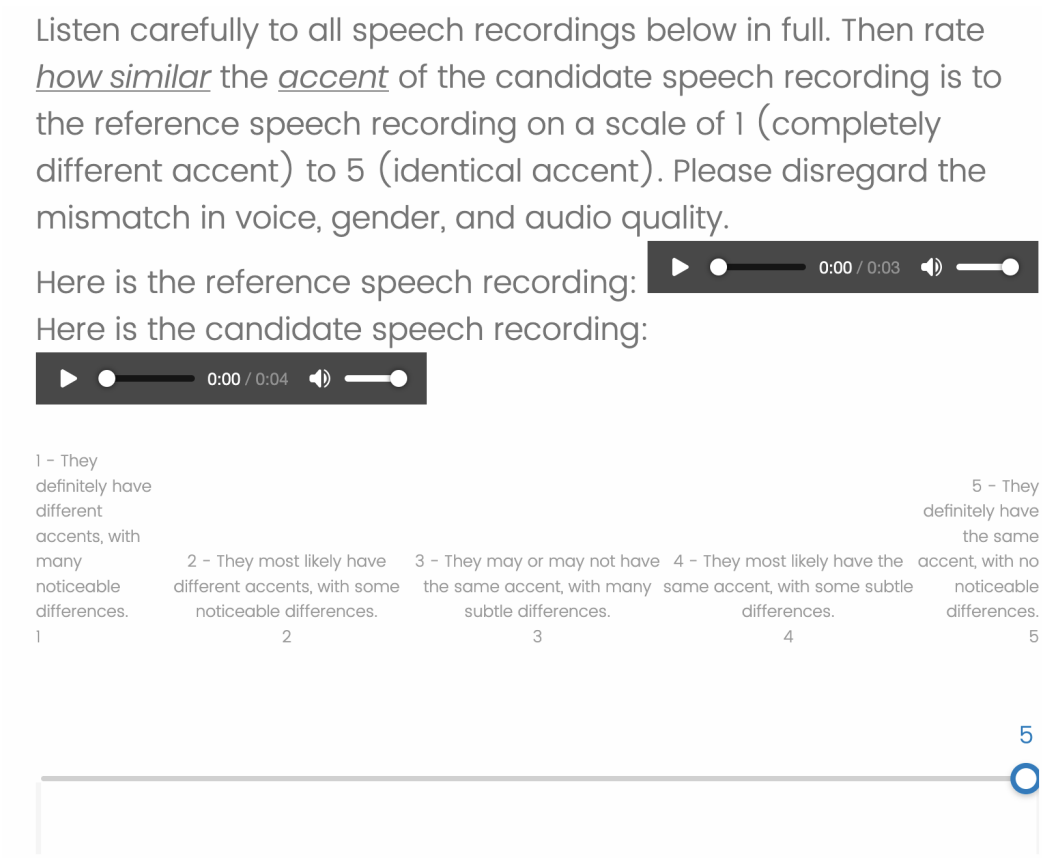
Malaysian

New Zealand

Appendix C

Appendix C

Presented here are example questions from the listening test described in Sections 4.1 and 4.4.


Listen carefully to all speech recordings below in full. Then rate how similar the accent of the candidate speech recording is to the reference speech recording on a scale of 1 (completely different accent) to 5 (identical accent). Please disregard the mismatch in voice, gender, and audio quality.
Here is the reference speech recording:
0:00 / 0:03
Here is the candidate speech recording:
0:00 / 0:04
1 - They definitely have different accents, with many noticeable differences.
1
2 - They most likely have different accents, with some noticeable differences.
2
3 - They may or may not have the same accent, with many subtle differences.
3
4 - They most likely have the same accent, with some subtle differences.
4
5 - They definitely have the same accent, with no noticeable differences.
5
5

Listen carefully to all speech recordings below in full. Then rate *how similar* the *emotion* of the candidate speech recording is to the reference speech recording on a scale of 1 (completely different emotion) to 5 (identical emotion). Please disregard the mismatch in voice, gender, and audio quality.

Here is the reference speech recording: 

Here is the candidate speech recording:

1 - They definitely display different emotions, with many noticeable differences.
1

2 - They most likely display different emotions, with some noticeable differences.
2

3 - They may or may not display the same emotion, with many subtle differences.
3

4 - They most likely display the same emotion, with some subtle differences.
4

5 - They definitely display the same emotion, with no noticeable differences.
5

3