Interspeech 2024
1-5 September 2024, Kos, Greece

Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges

Per E Kummervold, Javier de la Rosa, Freddy Wetjen, Rolv-Arild Braaten, Per Erik Solberg

National Library of Norway

per@capia.no, versae@nb.no, freddy.wetjen@nb.no, rolv.braaten@nb.no, pererik.solberg@gmail.com

Abstract

This paper presents NB-Whisper, a tailored adaptation of OpenAI's Whisper model, specifically fine-tuned to address the unique challenges of Norwegian language Automatic Speech Recognition (ASR). We highlight its key contributions and summarise the results achieved in converting spoken Norwegian into written forms and translating other languages into Norwegian. By training on a 22,000 hour weakly aligned dataset, we show that we are able to improve the Norwegian Bokmål transcription by OpenAI Whisper Large-v3 from a WER of 10.4 to 6.6 on the Fleurs Dataset and from 6.8 to 2.2 on the NST dataset.

Index Terms: speech recognition, language models, whisper

1. Introduction

Automatic Speech Recognition (ASR) holds the promise of changing the way we interact with technology by enabling machines to process human speech.

Early Norwegian speech recognition research initially focused on limited vocabularies suitable for telephone applications, confronting challenges like compound words and varied pronunciations of numbers (Svendsen et al., 1989; Paliwal, 1992). Funded by the European Union at the start of the century, subsequent projects expanded the scope to more complex linguistic elements, contributing to valuable datasets and technical advancements using hidden Markov models and Mel Frequency Cepstral Coefficients (Amdal and Ljøen, 1995; Höge et al., 1997). However, these systems, often utilising the Hidden Markov Model Toolkit (Young, 1994), were limited in handling open-ended recognition and struggled with out-of-vocabulary words or real conversations. The introduction of newer datasets in recent years has led to systems with improved performance, setting the stage for the Wav2Vec work described in De la Rosa et al. (2023) that seeks to address these longstanding challenges in Norwegian speech recognition.

OpenAI's Whisper (Radford et al., 2023) represents a significant deviation from the traditional ASR models. Unlike the standard approach of unsupervised pretraining followed by fine-tuning on a verbatim dataset mapping phonemes to graphemes, Whisper is pretrained on a vast, loosely aligned corpus of subtitles. This method has yielded impressive results, particularly in English, by not only handling capitalization and punctuation in a single step but also producing transcriptions that closely resemble written natural language.

However, Whisper's effectiveness diminishes when dealing with Norwegian, a language characterised by its rich dialectical diversity and two written standards: Bokmål and Nynorsk. To address this limitation, we developed NB-Whisper, a specialised adaptation that provides transcriptions for both Norwegian Bokmål and Nynorsk, as well as translations into English.

Our work aims to contribute to the general body of knowledge in ASR, providing strategies and frameworks that can be adapted for other languages, especially those with similar linguistic complexities. By pushing the boundaries in Norwegian ASR, we seek to pave the way for innovations and improvements in speech recognition technologies globally.

2. Language Variability and ASR Challenges

The Norwegian language encompasses two written standards with an equal official status: Bokmål and Nynorsk. Bokmål is close to the Oslo dialect and historically influenced by Danish. Nynorsk, written by about 13% of Norwegian pupils in 2023 (Statistics Norway, 2023), is close to a number of rural dialects, particularly on the west coast. The two written standards differ mostly in inflectional forms and vocabulary, but there are also some syntactic differences. In addition, the orthographic norm of each written standard allows for a significant degree of variation compared to other European languages, mirroring the variation found in dialects. For example, all feminine nouns in Bokmål such as "lua", 'the hat', can also be written in the masculine "luen". In Nynorsk, infinitives can end on -e or -a, as in "å elske" or "å elska", 'to love'.

The spoken dialects of Norwegian, while mutually intelligible, differ substantially in grammar, pronunciation and vocabulary. For example, the interrogative pronoun "who", written "hvem" in Bokmål and "kven" in Nynorsk, is recorded with 38 distinct pronunciations in the Nordic Dialect Corpus (Johannessen et al., 2012). There is no official norm of spoken Norwegian, and dialects are widely used, even in official settings such as in parliament and on the news.

ASR systems for Norwegian need to be able to handle the substantial dialect variation. Also, they should be able to transcribe in both written standards. Written transcripts such as meeting notes, subtitles, or parliamentary interventions, are usually expected to render the speech consistently in Nynorsk or Bokmål, regardless of the dialect of the speaker, so the written standard should not depend on the speaker's dialect. Moreover, ASR test sets need to contain transcriptions in both written standards, from speakers of different dialects, and with metadata about dialects in order to give a realistic picture of ASR performance.

3. Model Architecture and Training

Whisper utilises an encoder-decoder Transformer architecture, with up to 30 second audio chunks re-sampled to 16,000 Hz and

10.21437/Interspeech.2024-907

transformed into an 80-channel Mel spectrogram. For the latest Large-v3, this is expanded to a 128-channel Mel spectrogram (OpenAI, 2023). Feature normalisation and a two-layer convolutional encoder with GELU activation are applied, followed by sinusoidal position embeddings and transformer blocks. The decoder mirrors the encoder's structure, employing learned position embeddings. Byte-level BPE tokenization from GPT-2 is adapted for multilingual support (Radford et al., 2023).

We retain the core Whisper architecture to be able to initiate training from the released checkpoints. We also developed open-source training scripts for TPU-v4-pods, enabling dynamic data changes during training (The National Library of Norway, 2024), and improving the Jax implementation in the HuggingFace Transformers library.

We did tailor the training approach by adjusting specific hyperparameters and procedures to better suit Norwegian ASR needs. We increased the batch size to 1024 for all model sizes and scaled the learning rate accordingly. Warmup steps were slightly increased since we initialised from pretrained weights. Inspired by the Distil-Whisper work in Gandhi et al. (2023) and their regularization practices, we also adopted a reduced setting for the weight decay with regards to the original OpenAI's Whisper implementation. We found that adding BPE dropout (increased to 0.2) was beneficial. Lastly, we did not implement the stochastic dropout used in OpenAI's Large v2 and v3. Instead we used activation dropout as a substitute, noting some but limited effect.

Model	OpenAI Whisper	NB-Whisper
Tiny	1.5×10^{-3}	6×10^{-4}
Base	1×10^{-3}	4×10^{-4}
Small	5×10^{-4}	2×10^{-4}
Medium	2.5×10^{-4}	1×10^{-4}
Large	2×10^{-4}	7×10^{-5}

Table 1: *Learning rates.*

All our training starts from the original OpenAI multilingual checkpoints. For Tiny, Base, Small, and Medium, this means the training was initiated from the checkpoints released in September 2022, and converted to the HuggingFace formats a few weeks later. The Large checkpoints are however also updated to Large v2 in December 2022 and to Large v3 in November 2023. For training our Large model we start from the Large v3 checkpoint. However, we do also report the results from Large v1 and Large v2 in Table 4 for comparison. Our training comprises two phases: an initial 200,000-step training, followed by dataset cleaning (see Section 4), and a final 50,000-step training. The latter phase is done on a cleaner dataset, and the main purpose of this training is to reduce hallucinations.

4. Dataset

We train on the following sources:

(a) **NRK Subtitles**. Subtitles from the Norwegian Broadcasting Corporation (NRK), known for their non-verbatim style, present a challenge for verbatim speech recognition due to significant deviations from spoken words. The subtitles are aligned using the timestamps from the subtitles. The data contains consistent notation for separating simultaneous speakers, continued sentences across timestamps, whether it is a recording of simultaneous texting, as well as some rarer notation for denoting the name of the speaker, the language spoken, credit to staff, etc. We clean this to get only the spoken text and combine it into longer segments. Due to the non-verbatim nature of the transcripts, a word-to-word transcription would typically have a word error rate (WER) around 30% on this dataset. We extract audio segments without speech from the same recordings to prevent the model to output text when there is no voice.

(b) **Audio Books**. Professional narrators read these audio books with minimal deviations from the text. For aligning the text, we transcribed the text with NB-Wav2Vec2 1B[1] by De la Rosa et al. (2023), and used the ideas from Ljubešić et al. (2022) for the forced alignment. Due to the existence of multiple correct spellings for words in Norwegian, even the best possible WER score hovers around 1-2%.

(c) **The NST dataset**. This dataset was created by the now defunct Nordisk Språkteknologi. The dataset, which is distributed by the Language Bank at the National Library of Norway with an open licence, contains 540 hours of recordings of close to 1000 speakers with different dialectal backgrounds. As speakers tend to adapt their speech somewhat to the written language when they are reading out loud, the dataset has less evidence of dialectal grammar and vocabulary than datasets with spontaneous speech. (The Norwegian Language Bank, 2023).

(d) **The Stortinget Speech Corpus**. This is an open speech dataset created by the National Library of Norway in 2023 (Solberg et al., 2023). It consists of around 5,200 hours of transcribed speech from Stortinget, the Norwegian parliament. The transcriptions are extracted from the official proceedings of Stortinget using a string matching algorithm adapted from Ljubešić et al. (2022).

The NST dataset, the Stortinget dataset, and the NPSC (used for testing) are publicly available. We are releasing our cleaned versions of these datasets[2]. The release of the NRK dataset is still pending the establishing of the right legal framework. There are currently no plans for releasing the dataset based on the audio books. We are releasing all code used for training, including the data preprocessing and cleaning pipelines[3].

Our training process involves two stages. The first stage uses the complete dataset, and results in functioning models improving significantly on the OpenAI Whisper. However, after thorough testing, we still see issues with hallucinations. To improve on this we then use these models to run inference on the entire dataset. We identify various errors in the datasets, including misalignments and omissions. Instead of using a teacher-student approach where we train on the new output, we use the output to clean the dataset following several criteria:

- **Fuzzy Matching for First and Last Words**. We ensure at least a partial match (80% threshold) for the first and last words of each segment to filter out alignment errors.
- **Identifying Insertions**. We remove audio snippets where the target contains insertions not present in the audio, such as non-spoken transcriber comments. This is determined by finding target n-grams (longer than 3 words) that do not exist in any model predictions.
- **Detecting Omissions**. Snippets are deleted if they contain spoken phrases missing from the target text, identified by

[1] https://huggingface.co/NbAiLab/nb-wav2vec2-1b-bokmaal

[2] https://huggingface.co/NbAiLab

[3] https://github.com/NbAiLab/nostram

Hyperparameters	OpenAI Whisper	OpenAI Whisper Large v3	NB-Whisper
Updates	1,048,576	655,360[1]	200,000 + 50,000
Batch Size	256	1,024	1,024
Warmup Updates	2,048	2,048	10,000 / 5,000
Max grad norm	1	1	1
Optimizer	AdamW	AdamW	AdamW
β_1	0.9	0.9	0.9
β_2	0.98	0.98	0.98
ϵ	10^{-6}	10^{-6}	10^{-6}
Weight Decay	0.1	0.1	0.01
Weight Init	Gaussian Fan-In	Gaussian Fan-In	OpenAI Whisper
Learning Rate Schedule	Linear Decay	Linear Decay	Linear Decay
BPE Dropout	0	0.1	0.2
Stochastic Depth	0	0	0
Activation Dropout	0	0	0.1

Table 2: *Model training hyperparameters.*

Dataset	Stage 1 (hours)	Stage 2 (hours)
NRK - Subtitles	16,518	2,478
NRK - No caption	715	312
Audio Books	2,461	2,275
The NST Dataset	260	490
The Stortinget Speech Corpus	2,230	523
Total	**22,184**	**6,078**

Table 3: *Source datasets and their durations (in hours) across two stages. Additional data augmentations, such as translations and timestamps, are not included.*

the absence of common n-grams (longer than 3 words) in all model predictions.

- **NER Analysis with BERT**. Since names are often inserted or spelled out, especially in the parliament transcript, we perform Named Entity Recognition (NER) analysis using the NB-Bert base (Kummervold et al., 2021). We use the number of named entities as a filter.

Any snippet violating these filters is deleted from the training set. Our general approach is that we do not edit transcriptions. We only delete entries we do not have confidence in.

5. Experimental Setup and Evaluation

We evaluated our models using test sets from select datasets, specifically focusing on the Norwegian components:

- **Fleurs Dataset**: This Google-curated, out-of-domain dataset comprises read-aloud Norwegian Bokmål text snippets (Conneau et al., 2023). It provides a valuable context for assessing ASR model performance on unfamiliar content.
- **NST Dataset**: As detailed in Section 4, we used the test portion of this dataset distributed by The Norwegian Language Bank (2023). Care was taken to avoid speaker overlap between this test set and our training and validation data.

For evaluating model accuracy, we used the JiWER package, processing all texts to lowercase and removing punctuation prior to WER calculation. Our reporting of non-normalized scores aims to provide a clear and direct comparison across different datasets.

[1]Current values are for v2. It is reported that they have extended the dataset for v3 but to our knowledge, OpenAI has not yet disclosed the number of trained steps.

6. Results and Discussion

Evaluating the models was done both qualitatively and quantitatively. We found a close correlation between errors in the dataset and hallucinations in the final models. Implementing a two-step procedure for cleaning, where we used the initial models to clean the dataset, eliminated many of these issues.

Model	Bokmål		Nynorsk
	Fleurs	NST	Common Voice
NB-Wav2Vec2 1B	10.7	3.0	26.5
OpenAI Large v1	13.2	12.7	51.7
OpenAI Large v2	11.6	10.3	49.6
OpenAI Large v3	10.4	6.8	30.0
NB-Whisper Large (Ours)	**6.6**	**2.2**	**12.6**

Table 4: *Word error rates (WER) of the large Whisper and Wav2Vec2 models. Lower values indicate better performance. Best scores in bold. These models is current SOTA for Norwegian ASR.*

Model	OpenAI Whisper	NB-Whisper
Tiny (39M)	76.4	**15.2**
Base (74M)	56.8	**11.5**
Small (244M)	29.6	**8.3**
Medium (769M)	15.5	**7.2**
Large (1550M)	10.4	**6.6**

Table 5: *WER scores for Fleurs in Bokmål for the different model sizes. Lower values indicate better performance. Best scores in bold.*

Table 4 compares the performance of OpenAI and NB-Whisper models alongside NB-Wav2Vec2 models for both Bokmål and Nynorsk, highlighting their differences in archi-

Model	OpenAI Whisper	NB-Whisper
Tiny (39M)	73.7	**8.1**
Base (74M)	56.9	**4.9**
Small (244M)	27.2	**3.1**
Medium (769M)	14.6	**2.3**
Large (1550M)	6.8	**2.2**

Table 6: *WER scores for NST in Bokmål for the different model sizes. Lower values indicate better performance. Best scores in bold.*

Model	OpenAI Whisper	NB-Whisper
Tiny (39M)	>100	**28.0**
Base (74M)	>100	**23.2**
Small (244M)	>100	**19.9**
Medium (769M)	60.2	**17.0**
Large (1550M)	30.0	**12.6**

Table 7: *WER scores for Common Voice in Nynorsk for the different model sizes. Lower values indicate better performance. Best scores in bold. Note that WER scores might surpass 100 when the number of errors surpass the words in the reference. We see no point in reporting exact numbers in these cases.*

tecture, size and approach. Whisper models, designed for robust multilingual transcription and translation, effectively handle noisy or varied speech data through extensive labeled data and context understanding techniques. In contrast, Wav2Vec2 models Baevski et al. (2020) employ unsupervised pre-training from raw audio, followed by fine-tuning on labeled data. This makes them efficient for languages with scarce resources. However, the written output is based word by word on the input, and lacks certain features typical for written language like capitalization and punctuation. In comparison, the output from a Whisper model tends to more closely resemble written text in terms of quality.

Evaluating models quantitatively presents certain challenges. Our focus is on scripted and read aloud text, where model comparability is highest. However, issues arise in normalization, like standardizing outputs amid speech diversity and lacking a universal dataset, complicating direct comparisons. Number normalization is notably difficult due to varying expressions and language-specific rules. In this study we did choose a light normalisation, mainly on punctuation and capitalization[4]. Increasing the normalisation and standardisation, could have led to lower scores.

Tables 6 and 7 detail the performance of OpenAI Whisper and NB-Whisper models across various model sizes, comparing their effectiveness in recognizing Bokmål and Nynorsk languages. In the context of NST Bokmål dataset, the performance metrics indicate a clear trend: as the model size increases from Tiny (39M parameters) to Large (1550M parameters), the accuracy improves significantly for both OpenAI Whisper and NB-Whisper models, with NB-Whisper generally outperforming its OpenAI counterpart across all sizes. Specifically, the Tiny model shows a stark contrast in performance between the two, with NB-Whisper Tiny achieving a 8.1% accuracy compared to OpenAI's Whisper's 73.7%. This trend continues down to the Large model, where NB-Whisper achieves a 2.2% word error rate, followed by OpenAI Whisper at 6.8%.

[4]A stronger normalization was applied in De la Rosa et al. (2023) for the NB-Wav2Vec2 models (e.g., spelling out numbers), hence the difference in scores.

In the Common Voice Nynorsk evaluation, the discrepancy in performance is even more pronounced. However, NB-Whisper achieves a remarkable 28.0% WER score even at the Tiny model size. As the model size increases, performance improves dramatically, with the Large model achieving a 12.6% error rate NB-Whisper and 30.0% for OpenAI Whisper. This highlights the critical impact of model size on the accuracy of speech recognition tasks in different languages, particularly in less commonly supported languages like Nynorsk.

7. Challenges and Limitations

The model's handling of hallucinations and long text transcriptions was only assessed qualitatively. This highlights a need for comprehensive assessment tools in Norwegian ASR research. Developing such evaluation tools would be vital for further advancing Norwegian ASR.

The model's architecture, while optimised for transcribing 30-second audio clips, presents a notable limitation in live transcription scenarios. The reliance on an autoregressive decoder is not ideal for doing real-time transcription. This suggests that alternative architectures might be more apt for applications requiring continuous, live transcription.

8. Future Work

Future improvements can be made in dataset cleaning. Our multi-stage approach, involving interim models, inherently risks discarding valuable data due to the fine line between high-quality and faulty data. Re-evaluating the entire dataset with improved tools developed during the project promises enhancements in data quality. Norwegian's diverse spellings pose a challenge for consistent model outputs. Fine-tuning large language models to standardize the corpus can potentially improve ASR for languages with high orthographic variation.

9. Conclusion

The creation of NB-Whisper marks a step in advancing Automatic Speech Recognition (ASR) for Norwegian, adapting OpenAI's Whisper to meet the language's unique challenges. Our efforts in refining ASR capabilities have shown promising improvements in handling Norwegian's diverse dialects and written standards. However, we recognize that our project has limitations, especially in evaluating certain aspects like hallucinations and long text transcriptions, since there are not enough detailed Norwegian datasets to thoroughly test these areas.

Continuous research in ASR technology is essential. There is a clear pathway for further development, especially in model architecture and expanding linguistic adaptability. A particular focus should be on increasing the consistency in orthographic variation.

10. Funding and Acknowledgement

Our research greatly benefited from the support provided by Google's TPU Research Cloud (TRC), which generously supplied us with Cloud TPUs essential for our computational needs. We also extend our gratitude to Google Cloud for their support through Google credits. Special thanks are due to Sanchit Gandhi at HuggingFace for his substantial assistance in developing the TPU training scripts, a crucial component of our project. Thanks to Njaal Borch for providing the code enabling the initial alignment of the texts from NRK.

11. References

Ingunn Amdal and Harald Ljøen. Tabu.0 - en norsk telefontaledatabase. Technical Report 95, Scientific Report, 1995.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. pages 798–805, 01 2023. doi: 10.1109/SLT54892.2023.10023141.

Javier De la Rosa, Rolv-Arild Braaten, Per Egil Kummervold, Freddy Wetjen, and Svein Arne Brygfjeld. Boosting norwegian automatic speech recognition. pages 555–564, May 2023. URL https://aclanthology.org/2023.nodalida-1.55.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling, 2023.

Harald Höge, Herbert S. Tropf, Richard Winski, Henk van den Heuvel, Reinhold Häb-Umbach, and Khalid Choukri. European speech databases for telephone applications. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:1771–1774 vol.3, 1997. URL https://api.semanticscholar.org/CorpusID:7852982.

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Anders Nøklestad, and André Lynum. The Nordic dialect corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3387–3391, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/773_Paper.pdf.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.3.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In Darja Fišer, Maria Eskevich, Jakob Lenardič, and Franciska de Jong, editors, *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.parlaclarin-1.16.

OpenAI. Whisper large v3. https://huggingface.co/openai/whisper-large-v3, 2023. Accessed: Feb. 01, 2024.

Kuldip K. Paliwal. On the use of line spectral frequency parameters for speech recognition. *Digit. Signal Process.*, 2:80–87, 1992. URL https://api.semanticscholar.org/CorpusID:16553299.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Per Erik Solberg, Pierre Beauguitte, Per Egil Kummervold, and Freddy Wetjen. A large Norwegian dataset for weak supervision ASR. In Nikolai Ilinykh, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, and Joakim Nivre, editors, *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 48–52, Tórshavn, the Faroe Islands, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.resourceful-1.7.

Statistics Norway. Table 03743: Pupils in primary and lower secondary school, by official form of norwegian 2023. https://www.ssb.no/en/statbank/table/03743, 2023. Accessed: Feb. 01, 2024.

Torbjørn Svendsen, Kuldip K. Paliwal, Erik Harborg, and P. O. Husoy. An improved sub-word based speech recognizer. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '89, Glasgow, Scotland, May 23-26, 1989*, pages 108–111. IEEE, 1989. doi: 10.1109/ICASSP.1989.266375. URL https://doi.org/10.1109/ICASSP.1989.266375.

The National Library of Norway. Nostram repository. https://www.github.com/NbAiLab/nostram, 2024. Accessed: Feb. 01, 2024.

The Norwegian Language Bank. Nst norwegian asr database. https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-54/, 2023. Accessed: Feb. 01, 2024.

Steve Young. The htk hidden markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44, 01 1994.