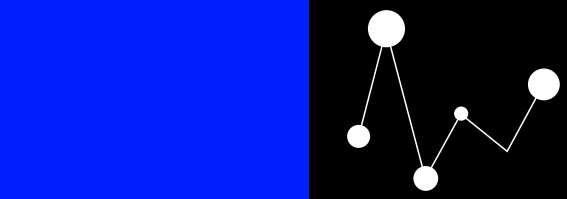
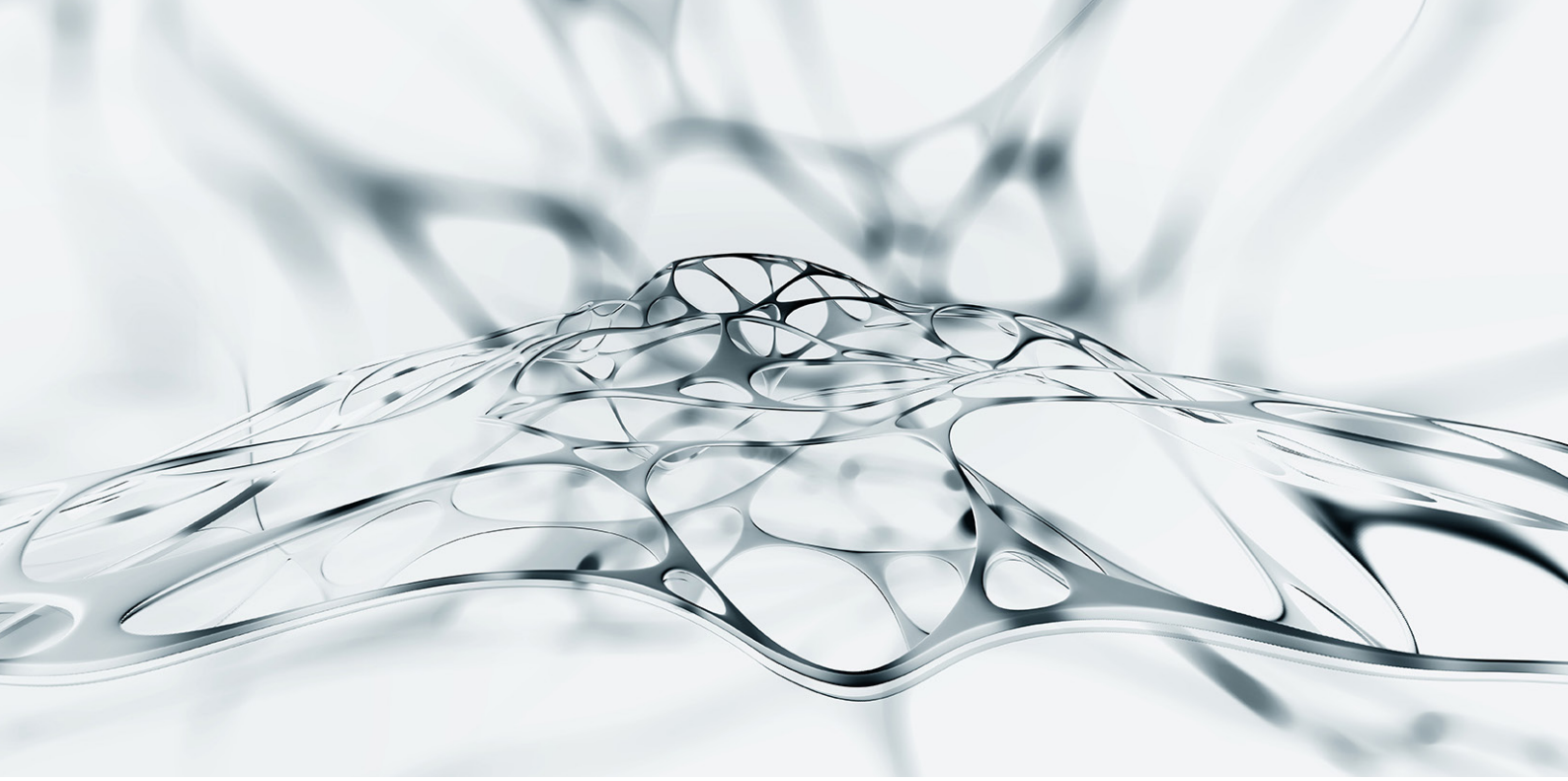
Enriching financial transcripts with automatic speaker labels

Lucy Evans
Applied ML Researcher, CESLT
London Stock Exchange Group



LSEG DATA & ANALYTICS



Contents

1. Executive summary

Speaker labelling technologies define a broad category of tasks that share a common goal of differentiating between speakers solely from their voice. The applications of these technologies are widespread. One example is the use of speaker identification in voice assistants, which enables the assistant to personalise its services to each individual user. Another is speaker verification, which is used in banking applications as an additional security measure. At London Stock Exchange Group (LSEG), we leverage the capabilities of speaker labelling technologies to assist with the production of transcripts for financial events and earnings calls. Each transcript is automatically labelled with a unique speaker identifier for every sentence – a process which, before the use of speaker diarisation, was a time consuming, manual task.

LSEG Analytics' speaker diarisation solution was developed by our in-house team of speech processing experts, the Centre of Expertise in Spoken Language Technologies (CESLT). The solution, which protects the privacy of the speakers involved, automates the insertion of anonymous speaker labels into the transcripts. Those labels are then manually corrected via a few simple clicks to assign real speaker names. With speaker labelling previously an entirely manual process, the implementation of speaker diarisation vastly improves the efficiency of the transcript production workflow. Put simply, the technology enables LSEG Analytics to produce speaker labels faster. With the labels being an essential component to a transcript, leveraging this technology allows for faster production of the transcripts, consequently enabling coverage across a wider range of financial events.

In this paper, we discuss the role that speaker diarisation plays in our best-in-class, proprietary transcripts production pipeline. We summarise the data science work undertaken prior to implementing this solution, including our considerations for respecting the privacy of the speakers involved. Finally, we discuss planned improvements to the current system, as well as the scope for future speaker analytics implementations using the outputs of the system. In summary, this paper showcases LSEG's applications of Natural Language Processing (NLP) within financial markets to explain the underlying analytics offered in our products.

2. Speaker labelling technology in the transcripts workflow

2.1 Use case

LSEG provides approximately 40,000 transcripts of financial events per year, which are produced by our team of highly skilled domain experts. The transcripts cover 10,000 companies, including all major global indices, for events such as Earnings calls, analyst meetings and corporate conference presentations[1]. Given the vast volume of events that are covered, and the rigorous standard of quality required, we consistently strive to develop innovative measures for increasing the efficiency and accuracy of the workflow used by our transcripts production team. One such development is the new automatic speaker labelling system, which we present in this paper. The system, developed using state-of-the-art speaker labelling technology, vastly reduces the time requirement for transcribers to add speaker labels to the transcripts.

The new transcripts production workflow, which includes the automatic speaker labelling system, is detailed in Figure 1. Automatic speech recognition (ASR) is first used to transcribe the speech in both recorded and live financial events. The speaker labelling solution then auto-populates the ASR transcript with anonymous speaker labels. Finally, the augmented transcript is edited into a final form by LSEG's domain expert transcribers, before being published via products such as LSEG Workspace. By bringing the transcripts generation in-house and enhancing the production workflow with ASR and speaker labelling technologies, LSEG Analytics has increased its coverage by over 2,700 companies, whilst also reducing costs and reducing our reliance on an external vendor. The capability provides high-quality output, which can further be enhanced and enriched with existing natural language processing capabilities developed by our Applied NLP team, such as summarisation, sentiment analysis and entity tagging.

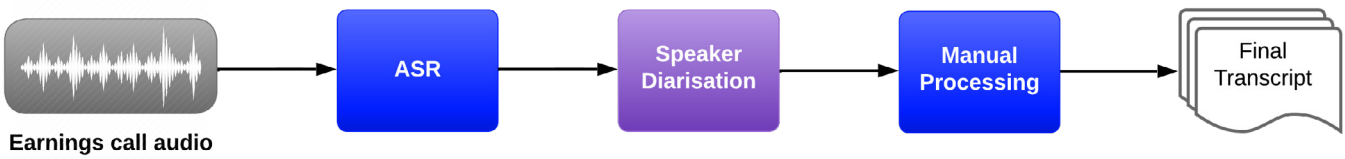


Figure 1: Transcripts production workflow
Source: LSEG, as of April 19, 2024

1 LSEG Transcripts & Briefs Database | LSEG

2.2 Selecting the appropriate technical solution

As a leading financial markets data provider, LSEG takes care to ensure that its analytics solutions are responsible, ethical and low risk. This includes our implementation of speaker labelling technology, which uses voice data to make predictions about speaker labels. Voice data is considered as personally identifiable information (PII), which LSEG protects in accordance with the General Data Protection Regulation (GDPR), as with all regulated data products. In order to manage the risk associated with processing sensitive data, a number of considerations were made prior to the development of the speaker labelling system.

Two technologies were considered for our speaker labelling solution: speaker identification (ID) and speaker diarisation. Speaker ID identifies and labels named speakers in a recording, whereas speaker diarisation differentiates between participant speakers, labelling the transcript with anonymous speaker labels (see Figure 2). Although speaker ID seems an obvious solution for a speaker labelling task, we highlight some concerns with implementing this technology at scale.

As noted, speaker labelling technologies use voice data, which is personally identifiable. Each speaker labelling technology uses the data in different ways, and thus the risk to an individual speaker varies by task. For speaker ID, a voice model is created for every speaker to be identified by the system, and this is stored in a large speaker database. Voice models are considered biometric, each being representative of a named individual's voice. Under the GDPR, biometric models are classified as 'special category' data and as such, the creation and storage of the speaker database requires greater care and enhanced justification compared to other forms of processing. Unlike speaker ID, speaker diarisation does not require the creation of a biometric speaker database, given that its output speaker labels are anonymous and only meaningful in relation to the other labels predicted for that recording (see Figure 2). With this in mind, and given LSEG's commitment to building responsible and ethical analytics solutions, we conclude that speaker diarisation is the more appropriate modelling framework for our automatic speaker labelling system. We summarise the details of this technology in the following section.

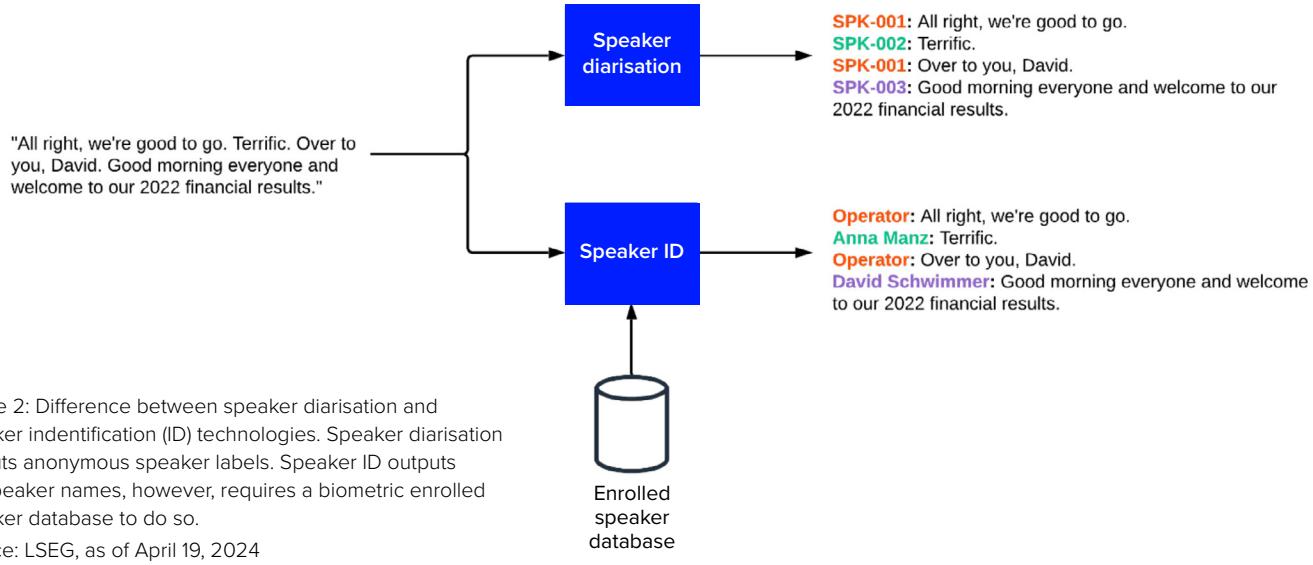


Figure 2: Difference between speaker diarisation and speaker indentification (ID) technologies. Speaker diarisation outputs anonymous speaker labels. Speaker ID outputs full speaker names, however, requires a biometric enrolled speaker database to do so.
Source: LSEG, as of April 19, 2024

2.3 Speaker diarisation

Speaker diarisation can be divided into three key stages: segmentation, speaker embedding extraction and clustering. The first stage involves segmenting an audio file into portions of speech and non-speech; often this involves the use of a speech activity detector. Any non-speech segments are discarded, as they are irrelevant to the speaker labelling task. Each speech segment is then passed to a speaker embedding model, which transforms the speech data into a quantified representation known as a speaker embedding.

In the first layer of the embedding model, the input speech segment is converted into a compressed representation of the signal. One popular representation in speech processing applications is the Mel Frequency Cepstral Coefficients (MFCCs), a compact representation of the audio that fits closely with the human auditory system. The speaker embedding model is trained to differentiate between different speakers using only these speech representations as input. As a result of this training paradigm, the later layers of the speaker embedding model encode the speech information that is most helpful in differentiating between different speakers. This information is compressed in a final embedding layer, from which speaker embeddings can be extracted.

A speaker embedding is extracted for every speech segment in the audio, and a clustering algorithm is then used to group the embeddings based on their similarities. The idea is that, because the embeddings encode information that is distinct to each speaker's voice, all embeddings from the same speaker will end up in the same cluster. Once the speaker clusters have been predicted, all their component embeddings are assigned the same anonymous speaker label. Each label is finally mapped back to the original audio segment represented by each speaker embedding. The process is summarised in Figure 3.

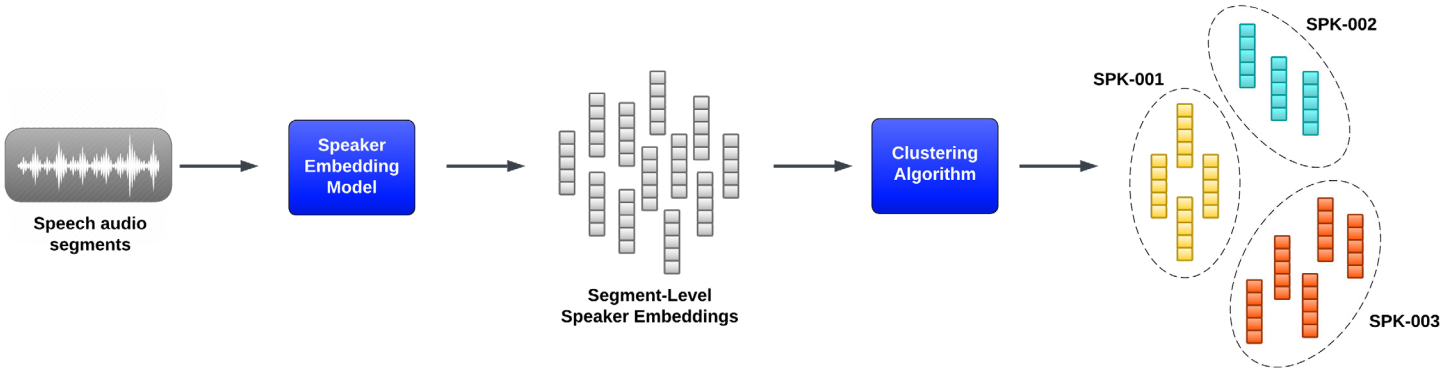


Figure 3: The process of Speaker Diarisation. A speaker embedding, which compresses useful information for differentiating between speakers, is extracted for each speech segment in an input recording. The embeddings are clustered into groups based on their similarities, and each group assigned a speaker label.

Source: LSEG, as of April 19, 2024

2.4 Speaker diarisation in the transcripts production workflow

LSEG Analytics' speaker diarisation solution works together with its ASR capability to maximise the efficiency of the transcripts production workflow. The integration of the ASR and speaker diarisation systems ensures that every sentence in an output ASR transcript is assigned a speaker label, and that there are no overlapping or conflicting segments between the two systems.

The predictions of the ASR model are first used to split a recording into speech and non-speech segments. Here, we make the assumption that the sentences predicted by the ASR system correspond to segments of speech in the recording. The ASR output transcript includes word-level timings, which are used to slice the full recording into sentence-level segments, based on the punctuation in the transcript. The portions of audio between the sentence segments are classed as non-speech and discarded. The speech segments are fed into the speaker diarisation system, where they are assigned anonymous speaker labels. The outputs of the ASR and speaker diarisation systems are integrated in HIVE, LSEG's internal transcripts application (see Figure 4).

The HIVE application is used for manual processing of the transcripts and enhances CESLT's speech technology outputs with additional functionality. With regards to the speaker labelling system, this includes the propagation of any speaker label change throughout the full transcript. For example, if the label "Unidentified_1" in Figure 4 were changed to "John Smith", HIVE automatically applies this change to every "Unidentified_1" instance in the transcript. The combination of the ASR and speaker diarisation technologies with LSEG's high-quality editing application provides a seamless production workflow, vastly reducing the time required for generating the final transcript. Specifically for the speaker labelling edits, the original task of an editor was to identify and label every speaker change point in the recording. By contrast, our speaker diarisation solution diminishes this task to a single correction of each pre-populated, anonymous speaker label to the speaker's real name.

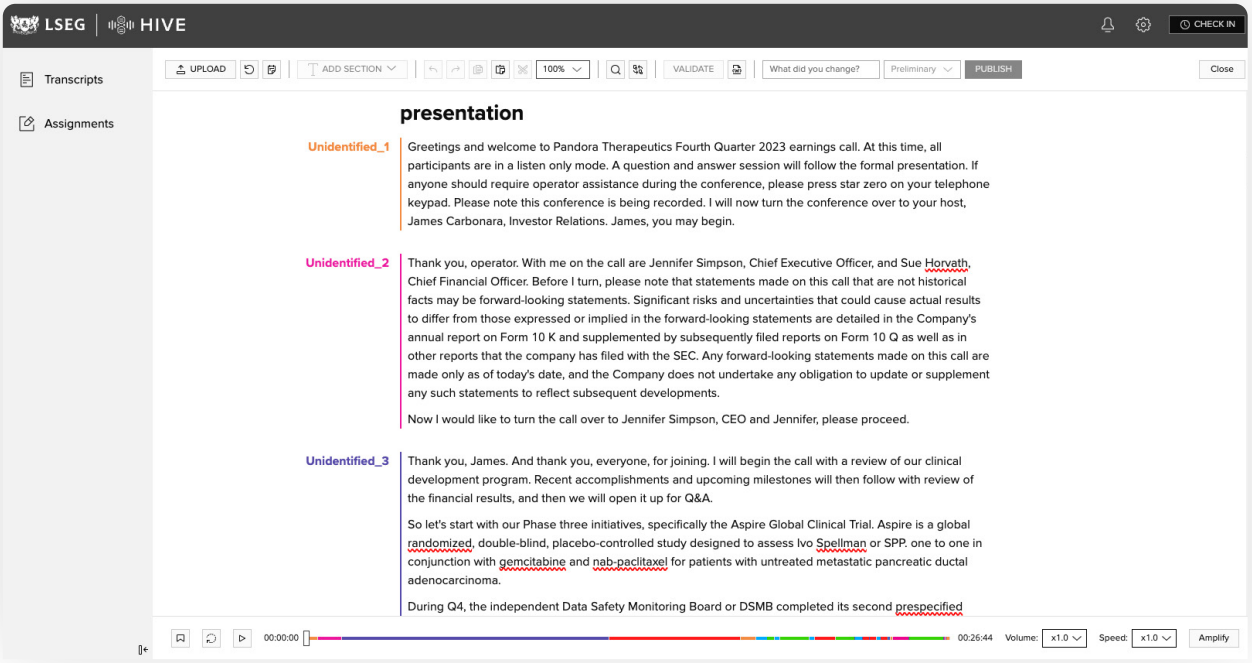


Figure 4: LSEG's internal transcripts application, HIVE. The application combines ASR and speaker diarisation outputs to display a speaker-labelled transcript, which is edited by domain experts prior to publishing on LSEG Workspace, etc.

Source: LSEG, as of April 19, 2024

3. Data science

In this section, we discuss CESLT's research effort into the automatic speaker labelling technology. Now with a clear focus on speaker diarisation, we summarise a series of experiments carried out to determine the best-performing speaker embedding model for our use-case. In our experiments, we investigate how models trained on proprietary, in-domain transcripts data perform on the speaker diarisation task compared to out-of-domain, pre-trained and publicly available speaker embedding models.

3.1 Data

The data required for training a speaker embedding model consists of segmented speech audio accompanied by a speaker label for every segment. The proprietary LSEG data set used, which we refer to as "the transcripts data", consists of historic financial event recordings and their accompanying transcripts. The transcripts include speaker names next to each of their contributions. This data set was produced prior to the implementation of LSEG Analytics' ASR solution.

We used CESLT's speech activity detection and forced alignment technologies to segment the audio into speech segments, each of which was labelled with its corresponding speaker label from the transcript. The transcripts were originally labelled with real speaker names. Speaker metadata, such as job title and corporation, was also provided for each speaker. For the purpose of data protection, all speaker names were anonymised via the creation of a unique and anonymous speaker hash for each speaker. The hashes were derived from the speaker metadata provided in each transcript but bear no relation to this in their final form. For model training, we selected a subset of the full transcripts data set that reflects the domain and distribution of data we would expect to see in production use of the system. The training set contains 33 different event types originating in 92 different countries; however, most recordings are from earnings calls and corporate conference presentations that took place in the United States. All calls take place in the English language and the full training set totals just over 3,000 hours of speech audio. The full count of distinct speakers in the data set, as calculated from the number of speaker hashes, is 17,550.

Our evaluation data set also reflects the domain and distribution of the data expected in production use of the system. A selection of 74 sessions, largely comprising US-based earnings calls, were selected, totalling around 1,100 hours of data. The sessions were manually segmented and labelled with anonymous speaker labels to provide a 'gold standard' speaker labelling example to evaluate the system's predictions against.

3.2 Experiments

In our experiments, we investigated the performance of various speaker embedding models on the task of speaker diarisation in the financial events domain. In this paper, we report results on four. Two of the models were trained on our in-domain transcripts data, and another two were pre-trained and available online. For all experiments, we use an x-vector speaker embedding model (Snyder et al., 2018) combined with a spectral clustering algorithm (Ng, Jordan and Weiss, 2002).

When training our models on the transcripts data, we used the Kaldi speech recognition toolkit. Kaldi is a large, open-source codebase that provides training and evaluation scripts, or 'recipes', for various speech processing tasks. Our first x-vector model was trained using the Callhome diarisation recipe[2]. This uses Kaldi for all model training stages and is our baseline x-vector model. The second selected training recipe[3] uses Kaldi for data processing but carries out the model training stage using Tensorflow. This recipe was built in an effort to facilitate the implementation of speech processing tasks in more generic toolkits, but has also been shown to outperform Kaldi-based implementations on the task of speaker verification (Zeinali et al., 2019).

We additionally investigated the performance of some pre-trained x-vector models. The first[4] (Snyder et al., 2018) was developed using conversational telephone data from the switchboard (Graff et al.; 1998, 1999, 2001, 2002, 2004) and NIST Speaker Recognition Evaluation (Przybocki and Martin, 2001) data sets, thus providing an example of a model trained on out-of-domain data. This model was trained using the same Kaldi recipe as our baseline transcripts model. A direct comparison can therefore be made between the two models, which differ only in their training data. The second pre-trained model offers another out-of-domain example, having been trained on celebrity interviews from the VoxCeleb data sets (Nagrani, Chung, and Zisserman, 2017; Chung, Nagrani and Zisserman, 2018). This is a Pyannote model, available from HuggingFace[5], which uses learnable SincNet features as its audio representations (Bredin et al., 2020). This is different to the other models, which extract MFCC features to represent the audio. While MFCCs were engineered for speech-related tasks based on human perception, they may not preserve all information that is useful for differentiating between speakers, such as pitch. SincNet features, instead, capture this information, and have been shown to perform better than MFCCs on the tasks of speaker identification and verification (Ravanelli and Bengio, 2018).

2 kaldi/egs/callhome_diarization/v2 at master · kaldi-asr/kaldi · GitHub
3 GitHub - BUTSpeechFIT/x-vector-kaldi-tf: Tensorflow implementation of x-vector topology on top of Kaldi recipe
4 Callhome Diarizaton Xvector Model 1a
5 pyannote/embedding · Hugging Face

3.3 Results

We used the transcripts evaluation data set to evaluate each model's performance on speaker diarisation in the financial events domain. The task was evaluated on the standard metric of diarisation error rate (DER), defined as follows:

$$DER = \frac{missed\ speech\ (s) + false\ alarm\ (s) + speaker\ confusion\ (s)}{time}$$

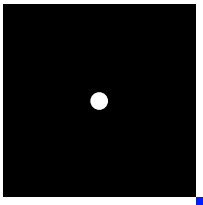
Here, missed speech refers to sections of speech audio which were falsely categorised as non-speech; false alarm refers to non-speech audio segments that were falsely categorised as speech; and speaker confusion refers to speech audio segments that were attributed to the wrong speaker. The metric therefore takes segmentation error, as well as speaker differentiation error, into account. A lower DER reflects a more accurate speaker diarisation system.

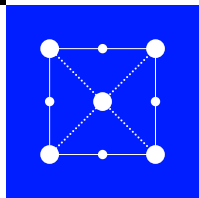
Due to the mismatch between training and test data domains, we expected that the pre-trained models would perform worse on this task than the in-domain models. However, as shown in Table 1, the out-of-domain models significantly outperformed those trained on the transcripts data. In particular, since the models have identical architectures and differ only in their training data, it is interesting to note the 6.2% reduction from the DER of the transcripts Callhome model (22.9%) to that of the pre-trained Callhome model (16.7%). This suggests a significant problem with the quality of the transcripts training data, which we discuss in the following section. A second result to highlight is that the Pyannote model, the only model to use SincNet features, achieves a substantially lower DER compared to all other models (12.6%). This further demonstrates Bredin et al. (2020)'s finding that the use of SincNet features may be favourable to that of MFCCs for the representation of audio data in speaker labelling tasks.

The results presented indicate that the Pyannote model is the most favourable model tested for our use-case of speaker diarisation in the financial events domain. The 12.6% error rate reflects 87.4% of transcript speaker labelling cases where the only manual edits required are a single edit of each predicted anonymous label to the speaker's real name. Furthermore, the HIVE application provides a simple interface for any edits required. Speaker labels are edited in two clicks – the first to select the speaker label in the transcript, and the second to select the desired, correct label. The combination of the speaker diarisation and HIVE technologies therefore provides a highly efficient, streamlined solution to speaker labelling in LSEG Analytics' transcripts production workflow. This vastly improves on the efficiency of the original, fully manual speaker labelling solution. Despite the performance of the current system being an impressive starting point, we continuously strive to improve the accuracy of the system in further research cycles, as described in section 4.2.

Model	Training data	Model details	DER (%)
Transcripts Callhome	In-domain Transcripts data	X-vector architecture using MFCC features.	22.9
Transcripts TensorFlow	In-domain Transcripts data	X-vector architecture, using MFCC features. Uses Kaldi for feature extraction and TensorFlow for model training.	19.6
Pre-trained Callhome	Telephone data from Switchboard and NIST SRE data sets	X-vector architecture using MFCC features.	16.7
Pre-trained Pyannote	Data from interviews uploaded to Youtube (VoxCeleb data sets)	X-vector architecture using SincNet features.	12.6

Table 1: Speaker diarisation error rate (DER) results





4. Discussion

4.1 Transcripts training data issues

The results from our experiments on the transcripts models revealed a significant issue with the structure of the transcripts training data, which we discuss in this section. Problems with training data are commonplace across machine learning applications, most commonly occurring when the data set was not specifically created for the purpose of model training. When data sets are created for this specific purpose, data set creators may have to go to extreme lengths to collect high-quality data that is appropriate for the intended use-case. An example of this within the speaker labelling field is the creation of VoxCeleb2 (Chung, Nagrani, and Zisserman, 2018), a large data set constructed for the purpose of speaker recognition applications. The group collected thousands of videos from Youtube, with the intention of labelling speech segments from the videos with real speaker names. The data set creation took place over seven stages, making use of numerous technologies such as face tracking, face verification and active speaker verification. The human resource, funding and technology required for this type of data collection is immense.

As mentioned in section 3.1, the transcripts training data set was obtained from historical financial recording transcripts, which were not produced with speaker embedding model training in mind. In this application, it is important that speaker labels in the training data set are accurate, and that all examples of speech from a single speaker are assigned the same speaker label. In section 3.1 we explained the pre-processing stage for anonymisation of the original speaker labels in the data set. Speaker hashes were generated directly from the speaker metadata provided in the original transcripts, such as name, job title and company. For this stage, we assumed that if the same speaker appeared in multiple recordings, their metadata would be the same for every recording, and that identical speaker hashes would therefore be generated for the speaker across all recordings. However, with the recordings spanning multiple years, we have identified significant variability in the metadata for the same speaker across recordings. For example, a single speaker in the data set may change role, company, or even name over the years, all of which result in a new hash being generated for that speaker. This may cause some speakers to be identified by two, or more, hashes in our training data set (see Figure 5).

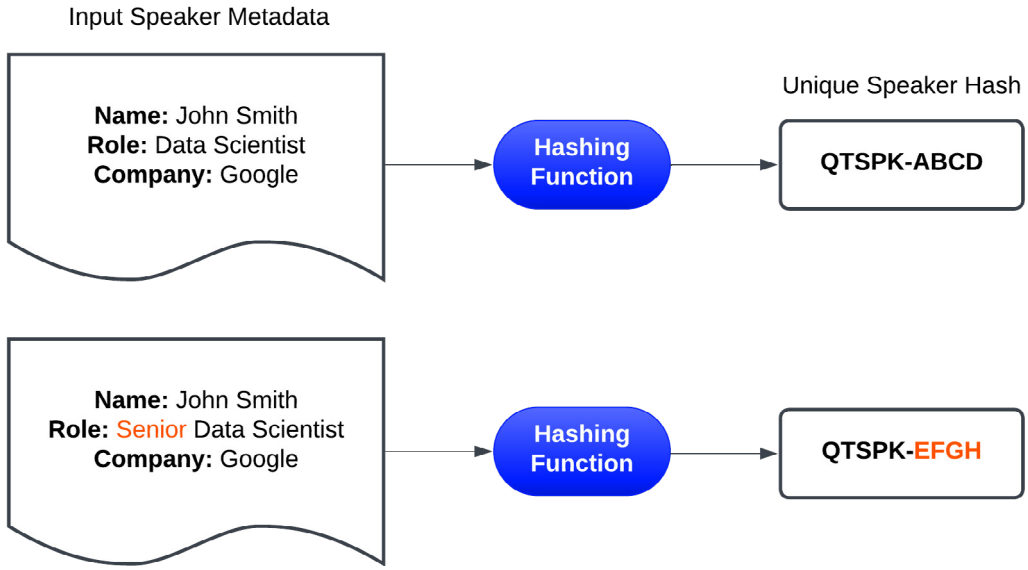


Figure 5: The speaker hashing system used, which results in a new speaker label for each unique instance of speaker metadata.
Source: LSEG, as of April 19, 2024

We consider this feature of the speaker labels in the transcripts data to be of sizeable importance for speaker embedding model training. During training, the model learns to discriminate between speech from all labelled speakers. The portions of speech which belong to the same speaker, but which are labelled with different hashes, become problematic as the system attempts to discriminate between the 'speakers' in those segments. The presence of multiple speaker hashes for the same speaker may force the system to focus on non-speaker variability, such as audio quality, background noise and so on, to differentiate between instances of the same speaker. This will subsequently affect the quality of the output speaker embeddings produced by the model, which in turn affects their performance on downstream tasks such as speaker diarisation. This quality issue in the embeddings is reflected in the DER results of the transcripts models.

4.2 Future directions

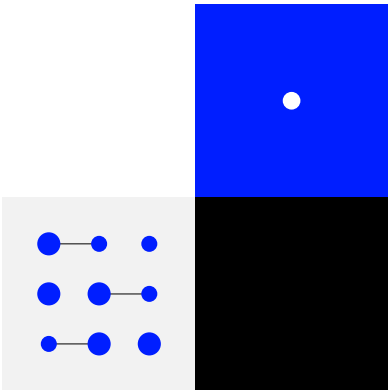
In this final section, we discuss improvements that we are investigating for the speaker diarisation system, as well as future use cases. A notable option for improving the performance of our domain-matched speaker embedding model is by fixing the speaker hashing issue in its training data set. However, this is a complex task; with 17,550 speaker hashes in the data set, a manual correction of the hashes is not feasible. We instead propose to address this issue using a distance algorithm, such as minimum edit distance, to identify similarities in cross-session speaker metadata. Using this algorithm, we will identify very similar instances of metadata, such as those in Figure 5, to predict instances that may belong to a single speaker. These predictions will be used to combine the resulting multiple speaker hashes into a single hash for that speaker, with the aim of minimising, if not fully eradicating, the issue.

If this data fix succeeds, we plan to use the training data to finetune a pre-trained model, since finetuned models tend to outperform models trained specifically on a single task, language or domain. This is likely due to the increased variability introduced by training on different data sets, which enables the model to generalise more widely, ultimately resulting in a more robust model. We propose to experiment with finetuning the Pyannote model, specifically, as well as experimenting with different speaker embedding model architectures, such as ECAPA-TDNN (Dawalatabad et al., 2021) and transformer-based models (Novoselov et al., 2022).

Looking to the future, we recognise that the speaker diarisation system has potential for a multitude of use cases beyond the transcripts production workflow. Speaker labels output by the system could be fed into downstream analytics tasks, such as to provide information about call engagement, the ability to search transcripts by speaker, and to skip to points in the transcript where a particular speaker has participated. Externally, the system also has scope for commercialisation, for example for providing a tool to tag external financial transcripts with speaker labels.

4.3 Conclusion

In this paper, we have discussed how speaker labelling technologies can be applied to real business scenarios to improve workflows. Specifically, we described the role that speaker diarisation plays in LSEG's transcripts production workflow. The integration of CESLT's speaker diarisation system with ASR technology and the HIVE application results in a highly efficient, innovative workflow, enabling LSEG to increase coverage and reduce delivery times of financial events transcripts. We are continuing to develop this workflow by investigating improvements to the accuracy of our training data speaker labels, experimenting with new speaker embedding model architectures, and developing further innovative features, such as downstream speaker analytics solutions.



References

Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W. and Gill, M.P., 2020, May. Pyannote audio: neural building blocks for speaker diarization. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7124-7128). IEEE.

Chung, J. S., Nagrani, A., and Zisserman, A. VoxCeleb2: Deep Speaker Recognition. INTERSPEECH, 2018.

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B. and Na, H., 2021. ECAPA-TDNN embeddings for speaker diarization. arXiv preprint arXiv:2104.01466.

Desplanques, B., Thienpondt, J. and Demuynck, K., 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Graff, D., Canavan, A., and Zipperlen, G. Switchboard-2 Phase I LDC98S75. Web Download. Philadelphia: Linguistic Data Consortium, 1998.

Graff, D., Walker, K., and Canavan, A. Switchboard-2 Phase II LDC99S79. Web Download. Philadelphia: Linguistic Data Consortium, 1999.

Graff, D., Walker. K, and. Miller, D. Switchboard Cellular Part 1 Audio LDC2001S13. Web Download. Philadelphia: Linguistic Data Consortium, 2001.

Graff, D., Miller, D., and Walker. K. Switchboard-2 Phase III Audio LDC2002S06. Web Download. Philadelphia: Linguistic Data Consortium, 2002.

Graff, D., Walker. K, and. Miller, D. Switchboard Cellular Part 2 Audio LDC2004S07. Web Download. Philadelphia: Linguistic Data Consortium, 2004.

Nagrani, A., Chung, J. S., and Zisserman, A. VoxCeleb: a large-scale speaker identification dataset. INTERSPEECH, 2017

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14 (pp. 849 – 856). MIT Press.

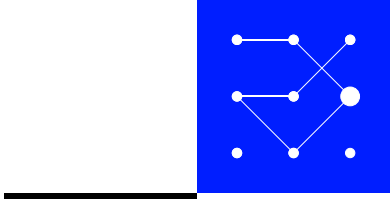
Novoselov, S., Lavrentyeva, G., Avdeeva, A., Volokhov, V. and Gusev, A., 2022. Robust speaker recognition with transformers using wav2vec 2.0. arXiv preprint arXiv:2203.15095

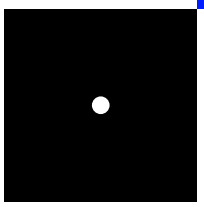
Przybocki, M., and Martin, A. 2000 NIST Speaker Recognition Evaluation LDC2001S97. Web Download. Philadelphia: Linguistic Data Consortium, 2001

Ravanelli, M. and Bengio, Y., 2018, December. Speaker recognition from raw waveform with sincnet. In 2018 IEEE spoken language technology workshop (SLT) (pp. 1021-1028). IEEE.

Snyder, D., Garcia-Romero, D. Sell, G. Povey D., and Khudanpur, S., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

Zeinali, H., Burget, L., Rohdin, J., Stafylakis, T. and Cernocky, J., 2019, May. How to improve your speaker embeddings extractor in generic toolkits. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6141-6145). IEEE.





Visit **lseg.com** | @LSEGplc LSEG

LSEG DATA & ANALYTICS

LDA3550959/10-24