<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr>
<td><table>
<tr><th>authors</th><td>- Ryant, N.
- Liberman, M.
- Yuan, J.</td></tr>
<tr><th>title</th><td>Speech activity detection on youtube using deep neural networks</td></tr>
<tr><th>publication_date</th><td>2013-08-25 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.CROSSREF</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td>10.21437/interspeech.2013-203</td></tr>
<tr><th>urls</th><td>- http://dx.doi.org/10.21437/interspeech.2013-203</td></tr>
<tr><th>id</th><td>id4589494756250490056</td></tr>
<tr><th>abstract</th><td></td></tr>
<tr><th>versions</th><td></td></tr>
</table></td>
<td><table>
<tr><th>authors</th><td>- Neville Ryant
- M. Liberman
- Jiahong Yuan</td></tr>
<tr><th>title</th><td>Speech activity detection on youtube using deep neural networks</td></tr>
<tr><th>publication_date</th><td>None</td></tr>
<tr><th>source</th><td>SupportedSources.SEMANTIC_SCHOLAR</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>- https://www.semanticscholar.org/paper/e651c1ec20460ae0f0fcd95f705370090a3bb335</td></tr>
<tr><th>id</th><td>id-8190058057386089388</td></tr>
<tr><th>abstract</th><td>Speech activity detection (SAD) is an important first step in speech processing. Commonly used methods (e.g., frame-level classification using gaussian mixture models (GMMs)) work well under stationary noise conditions, but do not generalize well to domains such as YouTube, where videos may exhibit a diverse range of environmental conditions. One solution is to augment the conventional cepstral features with additional, hand-engineered features (e.g., spectral flux, spectral centroid, multiband spectral entropies) which are robust to changes in environment and recording condition. An alternative approach, explored here, is to learn robust features during the course of training using an appropriate architecture such as deep neural networks (DNNs). In this paper we demonstrate that a DNN with input consisting of multiple frames of mel frequency cepstral coefficients (MFCCs) yields drastically lower frame-wise error rates (19.6%) on YouTube videos compared to a conventional GMM based system (40%).</td></tr>
<tr><th>versions</th><td></td></tr>
</table></td>
<td>DUPLICATES</td>
<td>188</td>
</tr>
</table>