6

EXPERIMENTOS Y RESULTADOS

En los capítulos anteriores se ha descrito los diferentes algoritmos que se utilizarán para realizar la tarea de *speaker diarisation*, y que en esta sección se emplearán de acuerdo al marco experimental que se describe a continuación.

Inicialmete, las pruebas consistieron en usar los algoritmos presentados para selección de modelo usando datos que fueron generados aleatoriamiente a partir de los parámetros de un HMM inicial; para tener una idea general de su desempeño individual.

Para estas primeras pruebas, se simuló una cadena de Márkov oculta con en base en parámetros fijos, con lo que se generó tanto una secuencia de datos observados, como los supuestos datos o variables ocultas que forman la cadena de Márkov. Se utilizó muestreo ancestral para la simulación de estos datos.

Para un caso en específico, se tiene lo siguiente. Realizando la inferencia de parámetros del HMM, se obtienen los siguientes resultados:

El primer algoritmo que se prueba, es el de selección de modelo usando un BIC.

Como ya se comentó, se usará una variante de BIC en donde se incorpora un término de regularización λ para que correspondan en órdenes de magnitud tanto la log-verosimilitud del modelo encontrado como su penalización respectiva.

El problema inmediato que se presenta, es cómo realizar la selección del parámetro de regularización λ que penalice de forma correcta la verosimilitud para los diferentes modelos propuestos. Si λ es demasiado pequeño, entonces la penalización realmente no tendrá efecto y dado el sobreajuste que se presenta al usar modelos más complejos, se preferirán siempre los modelos con más parámetros. Por otro lado, si al escoger λ se da demasiado peso al término de penalización, entonces siempre se preferirán los modelos más sencillos.

Para encontrar el valor de λ adecuado, se puede entonces formar una superficie con las diferentes curvas de selección BIC de acuerdo a cómo varía λ, e inspeccionar esta duperficie para encontrar una región de confianza en la que el valor de λ es el adecuado.

Por otro lado, para la segunda prueba, se procedió a usar bootstrap con la estadística log-likelihood ratio como ya se describió anteriormente en el **??**, y haciendo la prueba de hipótesis del modelo de n estados contra el de $n+1$ estados.

6.1 EXPERIMENTOS

Para los experimentos realizados, se generaron mediante un sintetizador de voz (también conocido como Text-To-Speech o TTS por sus siglas en inglés) que nos permitió tener un mayor control sobre el contenido como tal de las grabaciones, así como sobre los posibles ruidos o interferencias en las secuencias de audio.

Si bien, para probar el desepeño contra otras propuestas del estado del arte se suelen usar otro tipo de bases de datos, éstas suelen no estar disponibles de forma libre, por lo que preferimos generar nosotros un pequeño dataset con el sintetizador de voz.

Usando dos motores para el sintetizador de voz, uno con voces en inglés y otro con voces en español, se generaron 6 secuencias de audio (3 en cada idioma) cuya duración así como el número de interlocutores que participan varía.

6.1.1 *Secuencia 1: Edgar Allan Poe*

En esta primer secuencia se tomaron varios poemas del escritor Edgar Allan Poe, y se utilizaron 6 diferentes voces en inglés. La secuencia de audio original es de 12:06min.

Para la etapa de agrupación de los vectores MFCC con k-means++ se usaron 140 centros iniciales, mientras que el banco de filtros fue el mismo que anteriormente se mencionó.

[Biso6]

En las Figura 6.1 y Figura 6.2 se muestran los parámetros estimados para los diferentes modelos que se propusieron. La primer columna corresponde a los parámetros verdaderos, mientras que las demás son los parámetros obtenidos para algunos modelos.

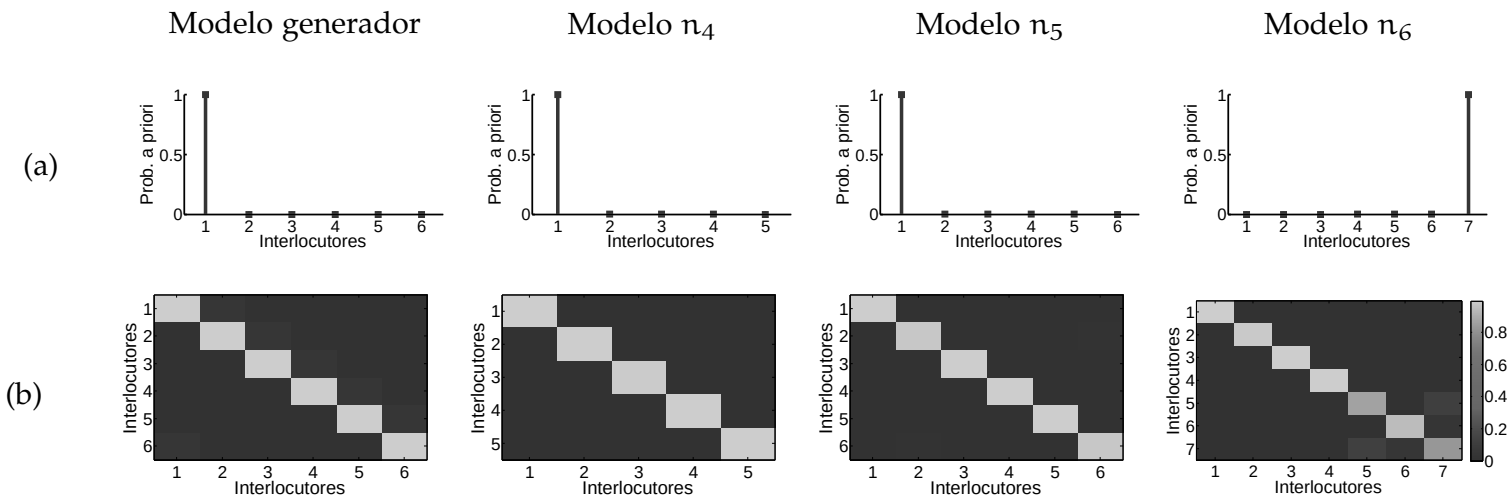

Figura 6.1: *Por columnas, se muestran los parámetros obtenidos para varios modelos propuestos. La primer columna corresponde a los parámetros verdaderos. En la primer fila* (a) *se muestran las probabilidades a priori de que cada uno de los interlocutores empiecen la conversación. En la segunda fila* (b) *se muestra en falso color la matriz de transición entre interlocutores para cada uno de los modelos mostrados.*

En la primer fila de la Figura 6.1 se muestran las probabilidades a priori de que cada uno de los interlocutores empiecen el diálogo. En la segunda fila se muestra en falso color la matriz de transición para cada uno de los modelos. Esta matriz representa la probabilidad de cambio entre las personas que participan en la conversación.

Se observa como en general para todos los modelos propuestos la matriz de transición que se recupera tiene una estructura diagonal, puesto que en este tipo de problemas, una persona suele hablar durante un periodo considerablemente largo, para que luego otra persona empiece a hablar.

En la Figura 6.2 se muestran las probabilidades de emisión para cada uno de los participantes, de acuerdo a las palabras en las que se ha discretizado la conversación, como se explicó anteriormente en el algoritmo (ref). Idealmente, cada interlocutor tiene asignadas con mayor probabilidad un cierto conjunto de palabras del diccionario, lo que hace que la identificación de las personas sea mucho más fácil de realizar.

Para la selección del modelo y siguiendo la metodología propuesta, primero se efectúa una inspección por medio de BIC regularizado para encontrar cuál o cuáles son los modelos más probables.

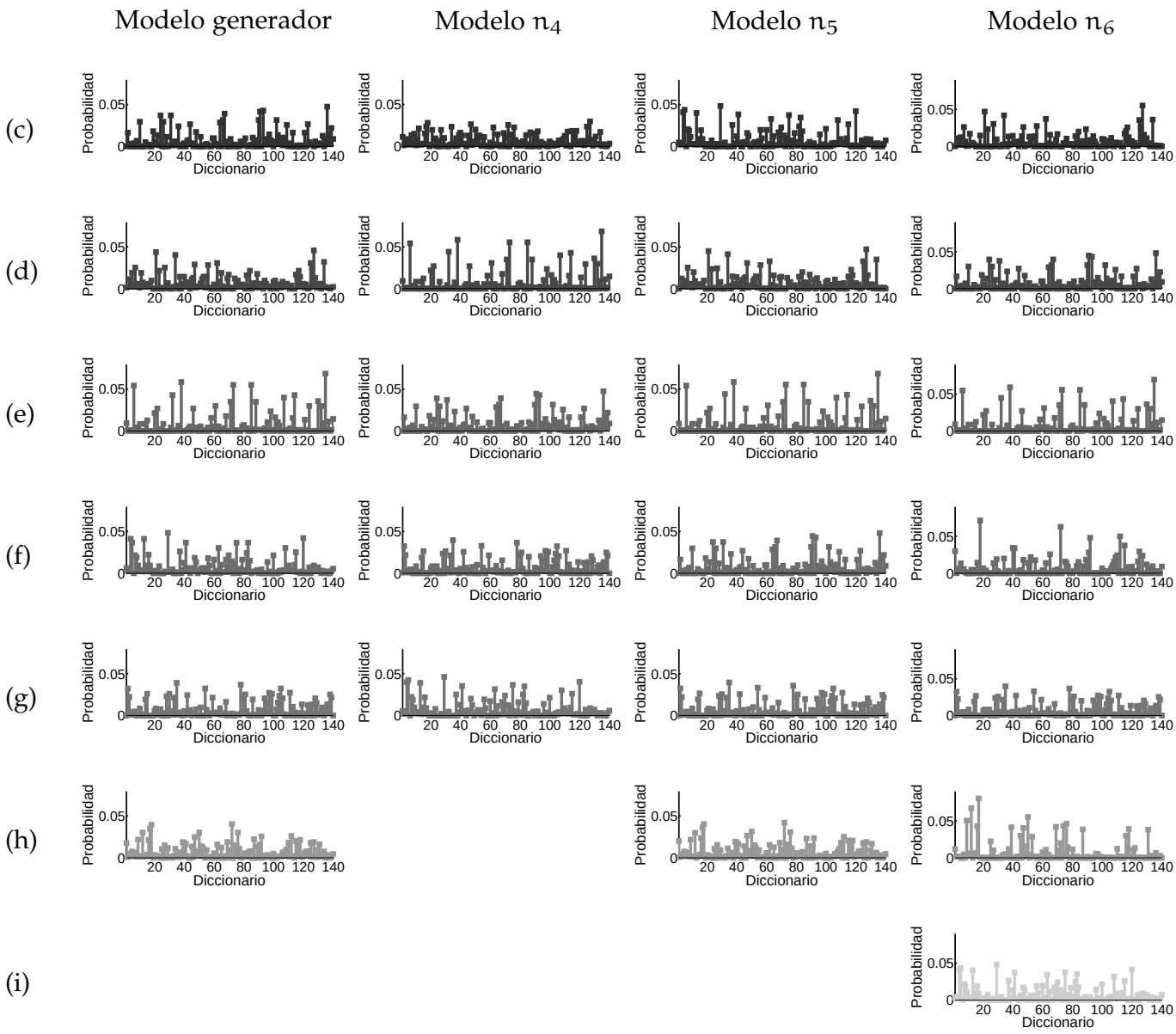

Figura 6.2: *Por columnas, se muestran las probabilidades de emisión para cada uno de los interlocutores de acuerdo al modelo. Cada fila representa las probabilidad de emisión de las palabras en el diccionario. La fila* (c) *para la primer persona, la fila* (d) *para la segunda persona y así de forma consecutiva, de acuerdo al número de personas que contempla cada modelo.*

En la Figura 6.3b se muestra la superficie generada al variar el valor de λ para diferentes curvas de selección BIC. Se observa cómo al principio λ es muy pequeño, y entonces el término de penalización no funciona por lo que se prefieren los modelos más complejos y sobreajustados. Por otro lado, cuando λ es muy grande, la penalización no permite mas que escoger los modelos más simples.

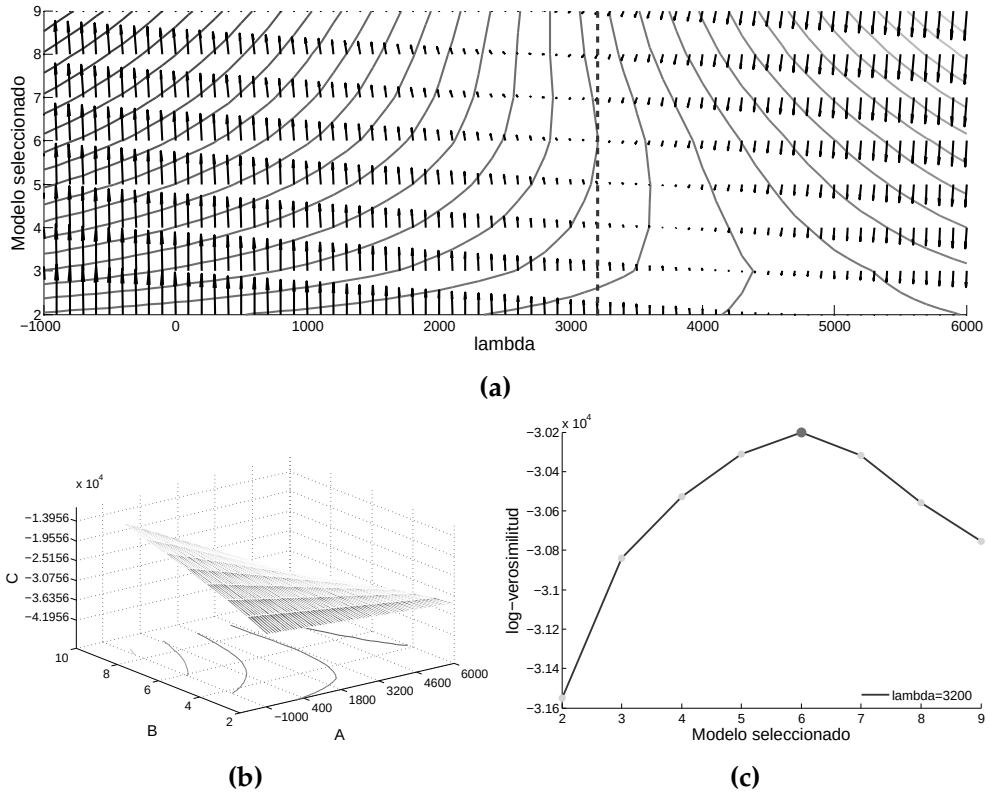

Figura 6.3: *En la Figura 6.3b se muestra la superficie generada al variar el valor de lambda para evaluar BIC. En la Figura 6.3a, se muestra las curvas de nivel de la superficie anterior, así como la dirección del gradiente en la misma. En la Figura 6.3c se muestra el valor seleccionado para lambda de acuerdo al analisis de sensibilidad realizado en la Figura 6.3a.*

Haciendo un análisis de sensibilidad se puede determinar cuál es el parámetro λ adecuado que penaliza de buena forma la log-verosimilitud.

En la Figura 6.3a se muestran las curvas de nivel de la figura anterior, además de la dirección del gradiente de la misma. Así mismo, en falso color se resaltan las zonas en las que el gradiente es menor.

Lo que nos interesa encontrar en la superficie, es el valor de λ que representa el punto de inflexión entre la selección de modelos demasiado complejos y modelos más simples. Para esto, se busca la zona en la que el gradiente sea lo más cercano a cero, pues implicaría que es un punto crítico.

Debido a la escala y a la forma en la que se calculó el gradiente, aunque para algunos valores cercanos de λ no haya mucha variación en esa dirección; si BIC está penalizando mal, entonces sí habrá gran variación para los diferentes modelos. Es por esto, que

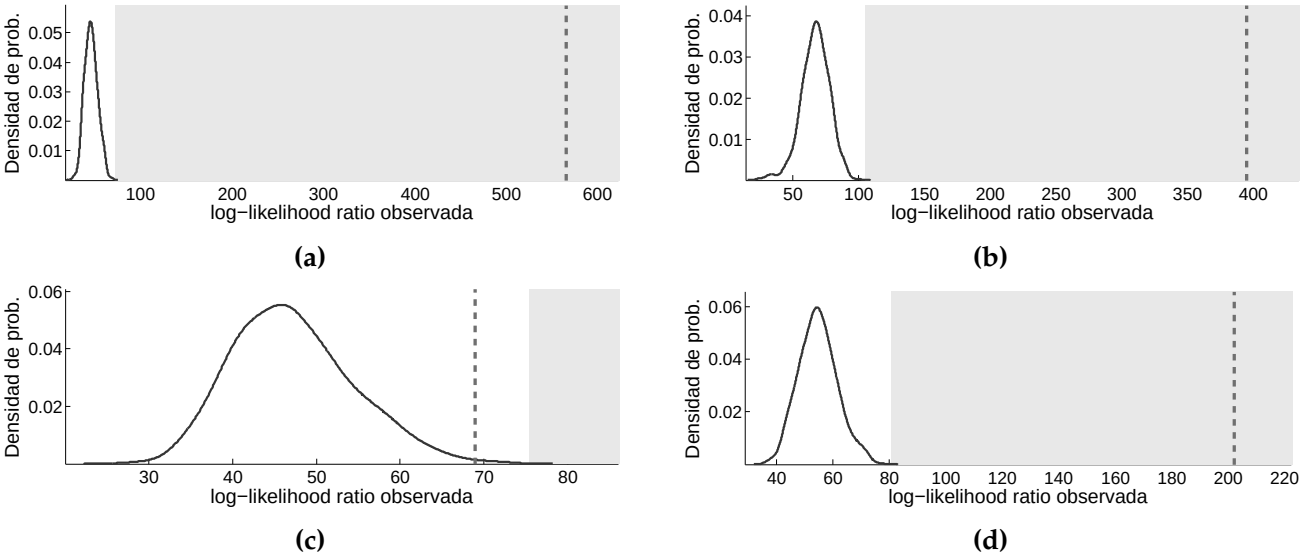

Figura 6.4: *En la Figura 6.4a se muestra la prueba de hipótesis realizada para comparar el modelo n_4 contra n_5. Como se observa, se rechaza la hipótesis de que el modelo correcto sea n_4. En la Figura 6.4b se hace la prueba del modelo n_5 contra n_6, y de la misma manera, se rechaza la hipótesis de que el modelo correcto sea n_5. Se sigue con la prueba de hipótesis del modelo n_6 contra n_7 en la Figura 6.4c, y en este caso el valor observado no cae dentro de la región de rechazo, por lo que no podemos rechazar que el modelo n_6 sea el correcto. Por último en la Figura 6.4d, se hace la prueba del modelo n_7 contra el modelo n_8, y se vuelve a rechazar.*

sólo en la zona en que la penalización sea del mismo orden de magnitud que la verosimilitud la variación en la curva BIC con el λ adecuado no será tan grande como en otras zonas.

Por último, se muestra en Figura 6.3c la curva BIC con el valor de λ encontrado a partir del análisis de sensibilidad realizado en la Figura 6.3a. El o los modelos que tengan un mayor valor BIC serán los que se seleccionarán como modelos ganadores.

En caso de que después de utilizar BIC se presente ambigüedad para determinar un modelo ganador, o ya sea para realizar un análisis más exhaustivo, se puede propone hacer una prueba de hipótesis para determinar cuál modelo se ajusta mejor a los datos.

A diferencia de la primer etapa, en la que se usa BIC como criterio para seleccionar el mejor modelo de un conjunto no definido de modelos con diferentes parámetros, la intención de hacer pruebas de hipótesis es determinar en un pequeño conjunto de probables modelos, cuál es mejor, y qué tan bueno es un modelo respecto a otro.

Al plantear la prueba de hipótesis se harán una gran cantidadad de simulaciones para ver qué tan bien se ajusta cada modelo a los datos originales, por lo que este proceso es computacionalmente intensivo y sólo se recomienda hacerlo para evitar la ambigüedad entre un par de modelos.

...

Por último, ya con el modelo seleccionado, se procede a calcular el error relativo de predicción que se obtuvo, comparando con el ground truth que se dispone para esa secuencia, con lo que se obtiene:

...

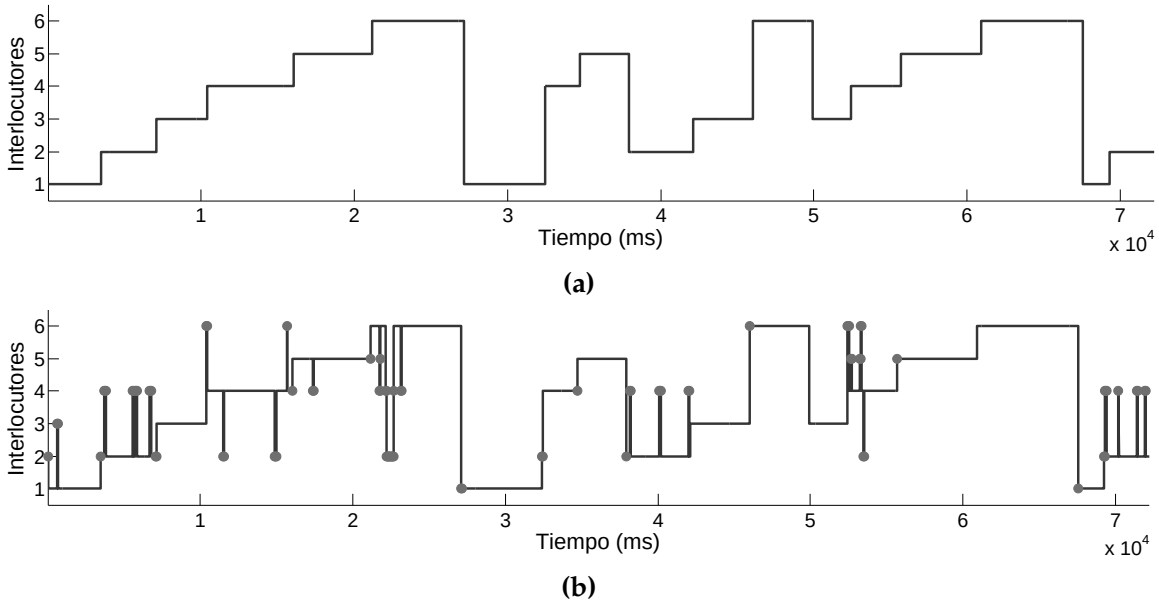

Figura 6.5: *La secuencia original en comparación con la secuencia recuperada del modelo ganador. En la segunda figura se muestran en rojo los errores cometidos respecto a la primera secuencia.*

Más a detalle, en la Figura 6.5 se observa en azul el orden en el que participan los interlocutores de acuerdo a la secuencia recuperada. En rojo se marcan tanto los falsos positivos como los falsos negativos, de acuerdo al ground truth. Hay que notar que cuando el número de estados para un modelo no es el correcto, entonces inminentemente el número de errores en la secuencia obtenida será mayor, pues al menos todas las intervenciones de un hablante no podrán ser emparejadas o serán asignadas a alguien más.

Se observa también que la mayoría de las veces, en la secuencia recuperada se encuentran algunos brincos entre personas, pero en esencia la estructura y el orden en que hablan los interlocutores es el correcto.

BIBLIOGRAFÍA

[AMBE+12] Xavier Anguera Miró, Simon Bozonnet, Nicholas W. D. Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech & Language Processing*, 20:356–370, 2012.

[AMWP06] Xavier Anguera Miró, Chuck Wooters, and José M. Pardo. Robust speaker diarization for meetings: Icsi rto6s meetings evaluation system. In Steve Renals, Samy Bengio, and Jonathan G. Fiscus, editors, *MLMI*, volume 4299, pages 346–358, 2006.

[BEF10] Simon Bozonnet, Nicholas W. D. Evans, and Corinne Fredouille. The LIA-EURECOM RT'09 Speaker Diarization System : enhancements in speaker modelling and cluster purification. In *ICASSP*, pages 4958–4961. IEEE, 2010.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[CH10] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[FBE09] Corinne Fredouille, Simon Bozonnet, and Nicholas W. D. Evans. The LIA-EURECOM RT'09 Speaker Diarization System. In *RT 2009, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, USA*, Melbourne, UNITED STATES, 05 2009.

[FE07] Corinne Fredouille and Nicholas W. D. Evans. The lia rt'07 speaker diarization system. In Rainer Stiefelhagen, Rachel Bowers, and Jonathan G. Fiscus, editors, *CLEAR*, volume 4625 of *Lecture Notes in Computer Science*, pages 520–532. Springer, 2007.

[FSJW11] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[IFM+05] Dan Istrate, Corinne Fredouille, Sylvain Meignier, Laurent Besacier, and Jean-François Bonastre. Nist rt'05s evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings. In Steve Renals and Samy Bengio, editors, *MLMI*, volume 3869 of *Lecture Notes in Computer Science*, pages 428–439. Springer, 2005.

[Jel98] Frederick Jelinek. *Statistical Methods for Speech Recognition*. Language, Speech, and Communication. MIT Press, 1998.

[MBI01] S. Meignier, J.-F. Bonastre, and S. Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. In *ISCA, A Speaker Odyssey, The Speaker Recognition Workshop*, Chiana (Crete), 18-22 Juin 2001 2001.

[MS91] David P. Morgan and Christopher L. Scofield. *Neural Networks and Speech Processing*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 1991.

[NLS09] Trung Hieu Nguyen, Haizhou Li, and Chng Eng Siong. Cluster criterion functions in spectral subspace and their applicationin speaker clustering. In *ICASSP*, pages 4085–4088, 2009.

[Rab89] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989. errata at http://alumni.media.mit.edu/ rahimi/rabiner/rabiner-errata/rabiner-errata.html.

[RJ93] Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[RS07] Lawrence R. Rabiner and R.W. Schafer. Introduction to digital speech processing. *Foundations and trends in signal processing*, 1:1–194, 2007.

[Ryd08] Tobias Ryden. EM versus Markov chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective. *Bayesian Analysis*, 2008.

[TR06] Sue Tranter and Douglas Reynolds. An overview of automatic speaker diarisation systems. *IEEE Transactions on Speech, Audio & Language Processing, Special Issue on Rich Transcription*, pages 1557–1565, 2006.

[WH07] Chuck Wooters and Marijn Huijbregts. The icsi rt07s speaker diarization system. In Rainer Stiefelhagen, Rachel Bowers, and Jonathan G. Fiscus, editors, *CLEAR*, volume 4625 of *Lecture Notes in Computer Science*, pages 509–519. Springer, 2007.