Speaker Diarization: A Top-Down Approach Using Syllabic Phonology

Erik Edwards[1], Amanda Robinson[1], Najmeh Sadoughi[1], Greg P. Finley[1], Maxim Korenevsky[1], Michael Brenndoerfer[2], Nico Axtmann[3], Mark Miller[1], and David Suendermann-Oeft[1]

[1] EMR.AI Inc., San Francisco, CA, USA
[2] University of California Berkeley, CA, USA
[3] DHBW, Karlsruhe, Germany
erik.edwards@emr.ai

Abstract. A top-down approach to speaker diarization is developed using a modified Baum-Welch algorithm. The HMM states combine phonemes according to structural positions under syllabic phonological theory. By nature of the structural phonology, there are at most 16 states, and the transition matrix is sparse, allowing efficient decoding to structural phones. This addresses the issue of phoneme specificity in speaker diarization – that speaker similarities/differences are confounded by phonetic similarities/differences. We address this here without the expensive use of a complete set of individual phonemes. The voice activity detection (VAD) issue is likewise addressed, giving a new approach to VAD.

Keywords: speaker diarization · speech activity detection · syllable

1 Introduction

When attempting the "who spoke when" question, i.e. speaker diarization, one must use features that distinguish different speakers of the dialog. These distinctions are confounded by phonemic differences, which are ultimately irrelevant to the labeling of speaker turns. This is the opposite of the situation in automatic speech recognition (ASR), where phone identities must be labeled, and speaker differences ignored. The problem in ASR is that of "speaker adaptation", whereas the problem in speaker diarization is sometimes referred to as "phoneme specificity" or "phone adaptive training". We present here a novel speaker diarization system that addresses the problem of phoneme specificity, while remaining highly computationally efficient.

The earliest approaches to diarization used a "bottom-up" approach of agglomerative clustering of feature vectors of different frames [14]. These are also called "unsupervised" in the sense that they require no labeled training data [35]. These approaches have remained heavily used in the literature [3,2]. Later systems began to introduce "top-down" approaches in combination with the bottom-up methods [37,40,12], but these require labeled training data. In fact, the first such paper [37] was also the first to introduce synthetic dialog data for

training purposes. Another early top-down approach [40] was the first to use HMM models with Baum-Welch training (although not as here, where we use it at diarization time). We first tried the bottom-up approach, where we found the issue of phoneme specificity to be strongly confounding. That is, for example, two fricatives from different speakers can be highly similar in their acoustic features, while a fricative and a vowel from the same speaker can be highly dissimilar. A number of papers have now addressed the problem of phoneme specificity/adaption in speaker diarization [18,5,39,43,4,31,36,42]. This issue is also well known in the larger literature on speaker recognition and verification [8,17]. We therefore abandoned the bottom-up approach in favor of the top-down approach presented here. This required a reliable set of training data, wherein both speaker labels and phone labels are available (since we desire to study phoneme specificity). Therefore, we also introduced our own synthetic corpus (Section 2, and described fully in the companion paper).

Our motivating application is the segmentation of doctor-patient dialogs, where the diarization is followed by ASR and information extraction [9]. Therefore, several of our basic decisions were guided by this application. First, the ASR stage requires MFCC features [7], so we attempt speaker diarization with the same MFCC features, but supplemented with a small number of auxiliary features. Second, we focus on the case of 2-speaker dialogs, which covers the great majority of doctor-patient encounters (although our approach is easily generalizable to 3+ speakers). Third, the issue of overlapped speech is less problematic in doctor-patient dialogs, because it is a situation where both members of the dialog have a high motivation to listen and to respect speaker turn taking. Other than yes/no responses, most medically critical information is delivered in longer turns with little or no overlap. Therefore, for our first system presented here, the focus is entirely on correct labeling of speaker identity, but not necessarily on refining the exact edges of speaker turns. In our system, each speaker-turn segment is submitted to the ASR stage with some leading/trailing audio anyway, so we have adopted the most typical "collar" used in diarization publications, which is 250 ms. The "collar" is a region around the segment boundaries that is ignored for computing the diarization error rate (DER) [1]. Finally, our system must operate in real-time, so there is a strong focus here on remaining computationally efficient at the time of diarization.

2 Synthetic diarization corpus

Doctor-patient dialogs are not freely available for diarization research. Existing data sets for diarization contain many speakers (e.g. meetings with 4 to 10+ speakers); or seem particular to a given situation or audio channel; or have speaker turns labeled, but not phonemic segmentations; or lack a large quantity of training data in addition to test sets; or cannot be obtained freely for general use. Therefore, we have developed a synthetic corpus as a basic starting point for diarization research, utilizing the open-source LibriSpeech corpus [27]. This synthetic corpus (Table 1) is described fully in the companion paper.

Table 1. Corpus of synthetic LibriSpeech dialogs

	Dialogs	Utts (Turns)	Tokens	Hours
Train	292	28522	989715	98.15
Dev-clean	48	2673	53765	4.98
Dev-other	45	2822	50227	4.69
Test-clean	43	2605	52279	5.07
Test-other	45	2861	51305	4.85

3 New lexicon with syllabic phonology

The concept of the syllable has a long tradition in linguistics, dating at least to the ancient Greek $\sigma\upsilon\lambda\lambda\alpha\beta\eta$ and Latin *syllaba* [38,25,16]. Use of the syllable in ASR dates to one of the earliest systems [26], and has recurred many times since [20,11]. However, syllabic approaches have consistently remained outside of the mainstream of ASR, and have been used only very rarely in speaker recognition [23,34,24]. We know of no syllable-based work in the speaker diarization or VAD literatures. One contributing factor may be the absence of a lexicon from which syllabic segmentations can be obtained directly. There is no simple method for obtaining syllabifications from ARPAbet-based lexicons [33], such as the widely-used CMUdict [28]. We have therefore developed an English lexicon utilizing syllabic phonology. For present purposes, this essentially means that each phoneme is assigned a structural position (i.e. Affix, Onset, Peak, Coda, Suffix), according to the most widely-accepted phonological theory [32,10,19,15].

The immediate practical motivation for introducing syllabic positions into our diarization work is that we would like to address phoneme specificity without however introducing a full phoneme-based decoding (as in ASR), which would be computationally expensive. On the other hand, there are only a handful of syllabic structural positions (5-15, depending on how many sub-positions are used), and the transition matrix for the structural positions is sparse. Thus, in the above 5-position scheme, Affix can only precede Onset; Onset can only precede Peak; Coda can only follow Peak; and Suffix can only follow Coda. An English utterance is a rather predictable succession of structural positions, and a dialog simply allows these to transition between speakers. Since the vowel phones occur exclusively in the Peak position, and since vowel segments are the dominant source of speaker characteristics, the Peak segments can be primarily used to distinguish speakers. This is the original idea and motivation; the resulting system in practice is given next.

4 Diarization method

Our speaker diarization system proceeds in two general stages: 1) Feature extraction and decorrelation/dimensionality reduction; 2) an expectation maximization (EM) algorithm to obtain posterior probabilities of HMM states, from which the speech/silence and speaker labels are obtained. All coding was done in C.

4.1 Feature extraction

Our total system cascades an ASR stage following diarization, so, for efficiency, we begin with the ASR acoustic features (40-D MFCCs [7]), supplemented with a small number of auxiliary features. Specifically, we append the 4-D Kaldi pitch features [13] and the 5-D VAD features of [29]. These are supplemented with Δ features, making a total 98-D feature set. This is reduced by PCA (principal component analysis) to 32-D output, followed by multi-class LDA (linear discriminant analysis) [41]. LDA was trained on labels defined by the 7 syllabic phone categories below, with vowels differentiated by the 251 unique speakers, giving 258 LDA labels total (1 silence, 6 consonant, and 251 vowel labels). All results presented here use a reduced set of 12-D LDA components. Finally, we change the 12-D LDA features to percentile units, where 128 bins were learned for each LDA feature from the training data. This allows the features to be held as char variables (the smallest data type in C), and used for direct table look-up, leading to greater computational efficiency at the time of diarization. Also, since the features are decorrelated by PCA/LDA, this allows the use of a direct (binned) probability representation, whereas GMM probability representations were found to perform worse and take $> 2\times$ longer computationally.

4.2 Modified Baum-Welch algorithm and HMM states

The Baum-Welch algorithm is a method to iteratively estimate the parameters of an HMM model [21]. As such, it is usually applied during training, and the resulting parameters fixed at decoding time. However, here we adapt the Baum-Welch algorithm to perform diarization on test data. The training data is only used to initialize the HMM parameters, and then the modified Baum-Welch algorithm adapts to the audio file under consideration by EM iterations. The update equations of the Baum-Welch are well-known and not covered here. More importantly, we have arrived at a method of progressive untying of HMM states with successive stages of iterations, such that stage 1 essentially provides a soft VAD output, and the last stage achieves the full diarization.

A recorded 2-person dialog consists of an initial segment of silence, alternating utterances of speakers 1 and 2 (with silent gaps within and between), and then a final segment of silence. The first person to speak is labeled "speaker 1" by definition, and "silence" includes any irrelevant background noise and often breath sounds. Note that initial silence is special in terms of the HMM A matrix, because the dialog must begin in this state, and this state must transition to speaker 1. However, we found no advantage to keep the final silence as a separate state, nor to keep within- vs. between-speaker silences separate. Thus, our HMM model has 4 overall states: 1) Speaker 1; 2) Speaker 2; 3) Initial silence; 4) Other silence. For the B matrix (emission probabilities), all silences remain tied together in one "tie-group".

Next, we split the Speaker 1 and 2 states according to syllabic phonology, in order to address phoneme specificity (see Introduction). The following split into 7 phoneme categories was found so far to perform best:

1. Prevocalic stops (B, D, G, K, P, T)
2. Prevocalic fricatives/affricates (CH, DH, F, HH, JH, S, SH, TH, V, Z, ZH)
3. Prevocalic liquids/nasals/semi-vowels (L, N, M, NG, R, W, Y)
4. Vowels (AA, AE, AH, ..., UW) (inclusive of all stress levels)
5. Postvocalic liquids/nasals/semi-vowels (L, N, M, NG, R, W, Y)
6. Postvocalic stops (B, D, G, K, P, T)
7. Postvocalic fricatives/affricates (CH, DH, F, HH, ..., Z, ZH).

This breakdown uses the most important phonemic distinction according to syllabic positions, which is the pre- vs. postvocalic distinction. This refers to consonants which lie before vs. after the vowel within the syllable. This distinction was emphasized already by Saussure (his"explosive" vs. "implosive" consonants) [30], and by the early Haskins studies of speech (their "initial" vs. "final" consonants) [6,22]. In terms of syllabic phonology, prevocalic merges Affix and Onset positions, postvocalic merges Coda and Suffix positions, and vowel is the same as Peak position. The pre- vs. postvocalic split was found to improve performance already at the VAD stage, whereas fewer distinctions (4 phone categories) and more refined distinctions (up to 15 phone categories) deteriorated performance. Thus, we proceed with the 7 structural-phone categories.

These phone categories define 7 HMM states per speaker, now giving 16 HMM states total (2 silence states + 7 states per speaker). Finally, we use the traditional 3 left-to-right substates per basic state, giving a grand total of $N = 48$ HMM states. Note that the major purpose of the 3 substates is to provide more realistic durational modeling by the transition matrix (A). For concreteness, we list these HMM states explicitly:

- HMM States 0-2: Initial silence
- HMM States 3-5: Other silence
- HMM States 6-8: Speaker 1, prevocalic stops
- HMM States 9-11: Speaker 1, prevocalic fricatives/affricates
- HMM States 12-14: Speaker 1, prevocalic liquids/nasals/semivowels
- HMM States 15-17: Speaker 1, vowels
- HMM States 18-20: Speaker 1, postvocalic liquids/nasals/semivowels
- HMM States 21-23: Speaker 1, postvocalic stops
- HMM States 24-26: Speaker 1, postvocalic fricatives/affricates
- HMM States 27-29: Speaker 2, prevocalic stops
- HMM States 30-32: Speaker 2, prevocalic fricatives/affricates
- HMM States 33-35: Speaker 2, prevocalic liquids/nasals/semivowels
- HMM States 36-38: Speaker 2, vowels
- HMM States 39-41: Speaker 2, postvocalic liquids/nasals/semivowels
- HMM States 42-44: Speaker 2, postvocalic stops
- HMM States 45-47: Speaker 2, postvocalic fricatives/affricates

The HMM A matrix, representing transition probabilities between these states, is learned once from the training data. Importantly, we do not update the A matrix during the modified Baum-Welch iterations. This is the most time-consuming update computation, and has negligible consequences for diarization.

Moreover, it was found that it was better to sparsify the A matrix by setting direct (0-ms lag) Speaker 1 to 2 transitions to 0.

The HMM B matrices, representing emission probabilities for each state, are first learned from the training data, and then updated with each iteration of the Baum-Welch during diarization. However, it is common practice to tie HMM states so that their emission probabilities are estimated jointly. This is particularly important if there is too little data. Moreover, most diarization systems begin with a VAD stage (speech vs. silence), before making the more refined distinctions for diarization. An important result of our preliminary investigations was that the B matrices are best updated with strong ties across states initially, and then progressive untying of the states towards the final diarization. We arrived at a 3-stage procedure, wherein the first stage uses only 7 tie groups, the last stage leaves most states untied, and the middle stage uses an intermediate degree of tying. Specifically, using the 48 HMM states enumerated above, the following 3-stages of state tie groups was found to work best:

STAGE 1 TYING OF B MATRIX:

- TIE-GROUP 0 == HMM States 0-5 (Silence)
- TIE-GROUP 1 == HMM States 6-8, 27-29 (Prevocalic stops)
- TIE-GROUP 2 == HMM States 9-11, 30-32 (Prevocalic fricatives)
- TIE-GROUP 3 == HMM States 12-14, 33-35 (Prevocalic liquids/nasals)
- TIE-GROUP 4 == HMM States 15-17, 36-38 (Vowels)
- TIE-GROUP 5 == HMM States 18-19, 39-41 (Postvocalic liquids/nasals)
- TIE-GROUP 6 == HMM States 20-22, 42-44 (Postvocalic stops)
- TIE-GROUP 7 == HMM States 23-25, 45-47 (Postvocalic fricatives)

It can be seen that no distinction is made in Stage 1 between speakers. This is therefore a speech vs. silence stage, except that speech has been expanded into the 7 structural-phone categories. This is, in fact, a new method of VAD, with soft (posterior probability) outputs. These are then used to initialize Stage 2 of the Baum-Welch iterations, where only the vowels are used to begin the separation of speakers. Thus, TIE-GROUP 4 of Stage 1 is split into 2 tie-groups in Stage 2.

STAGE 2 TYING OF B MATRIX:

- TIE-GROUP 0 == HMM States 0-5 (Silence)
- TIE-GROUP 1 == HMM States 6-8, 27-29 (Prevocalic stops)
- TIE-GROUP 2 == HMM States 9-11, 30-32 (Prevocalic fricatives)
- TIE-GROUP 3 == HMM States 12-14, 33-35 (Prevocalic liquids/nasals)
- TIE-GROUP 4 == HMM States 15-17 (Speaker 1 Vowels)
- TIE-GROUP 5 == HMM States 36-38 (Speaker 2 Vowels)
- TIE-GROUP 6 == HMM States 18-19, 39-41 (Postvocalic liquids/nasals)
- TIE-GROUP 7 == HMM States 20-22, 42-44 (Postvocalic stops)
- TIE-GROUP 8 == HMM States 23-25, 45-47 (Postvocalic fricatives)

It should be kept in mind that speaker distinctions are most usefully obtained from vowels. A major purpose of the consonant categories is just to separate them out from the vowels, so as not to contaminate the acoustic evidence provided during vowels states. Consonants also provide some degree of power to distinguish speakers, but we leave these states tied across speakers until the final iterations, in order not to interfere. Experiments showed that all 3 of these stages (and in this order of course-to-refined) were necessary to achieve the best performance. 8 EM iterations per stage were used for all results here.

Following the 24 EM iterations of the 3-stage Baum-Welch algorithm, the posterior probabilities are summed across all Speaker 1 states, all Speaker 2 states, and all Silence states. By this method, it is not important if the algorithm has perfectly separated various consonant categories, because they are all summed together with the vowel states for each Speaker. The final diarization label is taken as the maximum of these three probabilities for each time frame.

5 Results and Discussion

We present results for the synthetic LibriSpeech dialog corpus (Section 2), and for 2 recordings of doctor-actor dialogs. In the latter, a real doctor interviewed an actor patient (to avoid privacy issues). The doctors were male, and the patients female. Audio was recorded by a cell phone. The 2 dialogs were 6.4 min and 5.7 min in duration, and used for test data only. All training to initialize the HMM A and B matrices was done on the synthetic corpus.

For the synthetic LibriSpeech corpus, we obtain the following DERs, using a collar of 250 ms, as assessed with the widely-used md-eval-v21.pl script (from NIST). The same collar and script was used to asses the VAD error rate (VER).

Table 2. Results for synthetic LibriSpeech dialogs

	Mean DER	Max DER	Mean VER	Max VER
Dev-clean	0.66%	2.44%	0.62%	2.38%
Dev-other	0.94%	3.75%	0.90%	3.75%
Test-clean	0.95%	4.45%	0.78%	4.44%
Test-other	1.18%	5.58%	1.12%	5.42%

It can be seen that, using the liberal collar of 250 ms, the algorithm can successfully detect speech (VAD) and then diarize all of the development and test files. It must be emphasized that this is by no means a guaranteed result, and previous versions of our diarization methods obtained mean DERs closer to 5-10%, or worse (i.e., early bottom-up method). Also, the present algorithm under different settings would often fail on a small subset of files, e.g. obtain max DERs worse than 20-30%. The influential settings are: inclusion of VAD and pitch features; number of LDA components; types of phonological distinctions;

type of probability model for B matrices (e.g. GMMs performed worse); and, critically, the tying and progressive untying of HMM states during successive stages of the EM iterations.

Interestingly, the majority of the observed DER is due to VER (VAD error). Thus, the grand-mean DER was 0.93%, and the grand-mean VER was 0.85%, and it was common (under the liberal collar of 250 ms) to observe files with the same DER as VER, meaning that the algorithm rarely struggles to separate speaker characteristics, if the stage-1 (soft VAD) outputs are accurate. In fact, some of the VAD errors obtained may be considered spurious, as breath noise is not consistently treated in the forced alignments. The results imply that future improvements should first focus on the Stage 1 VAD phase.

For the live doctor-actor dialog recordings, we obtain:

Table 3. Results for recorded doctor-actor dialogs

	DER	VER
Dialog 1	4.06%	3.26%
Dialog 2	10.00%	9.13%
Average	7.20%	6.37%

Thus, a reasonable diarization of the real-world recordings was still obtained, despite the fact that the HMM model was trained only on synthetic data with no overlap. The LibriSpeech corpus is primarily American speech, wheras the doctor-actor dialogs here were British speech; and the recording method (cell phone) was quite different than for the training corpus. Also, the real-world dialogs contain many segments of coughing and other non-speech sounds that are not present in the training data, as well as many hesitation sounds ("umm", "ahh"). Finally, the manual diarization of these dialogs is likely not perfect. Therefore, the average DER of 7.2% is encouraging for the applicability of the general methods reported here, although we will clearly need to obtain matched training data for the methods to fully work.

6 Summary and Conclusion

We have presented our initial speaker diarization system, with the intended application of doctor-patient dialogs. Training on a synthetic corpus, to initialize HMM parameters, allowed successful diarization of recorded doctor-patient dialogs. The HMM parameters are updated in 3 stages of EM iterations, at the time of diarization. Emphasis was on computational efficiency, leading to a reduced Baum-Welch algorithm that omits A-matrix updates, and uses discrete (binned) probability distributions. HMM states are based on only 7 structural phones, as motivated by syllabic phonological theory, with sparse transition matrix, allowing an efficient approach to the phoneme specificity problem. The first of the 3 EM stages replaces the usual VAD stage, also improving total efficiency.

References

1. Anguera Miró, X.: Robust speaker diarization for meetings. Ph.D. thesis, Univ. Politècnica de Catalunya (2006)
2. Anguera Miró, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process **20**(2), 356–370 (2012)
3. Anguera Miró, X., Wooters, C., Peskin, B., Aguiló, M.: Robust speaker segmentation for meetings: the ICSI-SRI Spring 2005 diarization system. Lect Notes Comput Sci **3869**, 402–414 (2006)
4. Bozonnet, S., Vipperla, R., Evans, N.: Phone adaptive training for speaker diarization. In: Proc INTERSPEECH. pp. 494–497. ISCA (2012)
5. Chen, I.F., Cheng, S.S., Wang, H.M.: Phonetic subspace mixture model for speaker diarization. In: Proc INTERSPEECH. pp. 2298–2301. ISCA (2010)
6. Cooper, F., Delattre, P., Liberman, A., Borst, J., Gerstman, L.: Some experiments on the perception of synthetic speech sounds. J Acoust Soc Am **24**(6), 597–606 (1952)
7. Edwards, E., Salloum, W., Finley, G., Fone, J., Cardiff, G., Miller, M., Suendermann-Oeft, D.: Medical speech recognition: reaching parity with humans. In: Proc SPECOM. vol. LNCS 10458, pp. 512–524. Springer (2017)
8. Fakotakis, N., Tsopanoglou, A., Kokkinakis, G.: A text-independent speaker recognition system based on vowel spotting. Speech Commun **12**(1), 57–68 (1993)
9. Finley, G., Edwards, E., Robinson, A., Sadoughi, N., Fone, J., Miller, M., Suendermann-Oeft, D.: An automated medical scribe for documenting clinical encounters. In: Proc NAACL. ACL (2018)
10. Fudge, E.: Branching structure within the syllable. J Linguist **23**(2), 359–377 (1987)
11. Fujimura, O.: Syllable as a unit of speech recognition. IEEE Trans Acoust **23**(1), 82–87 (1975)
12. Gauvain, J.L., Adda, G., Lamel, L., Adda-Decker, M.: Transcribing broadcast news: the LIMSI Nov96 Hub4 system. In: Proc DARPA Speech Recognition Workshop. pp. 56–63. DARPA (1997)
13. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: Proc ICASSP. pp. 2494–2498. IEEE (2014)
14. Gish, H., Siu, M.H., Rohlicek, J.: Segregation of speakers for speech recognition and speaker identification. In: Proc ICASSP. vol. 2, pp. 873–876. IEEE (1991)
15. Goldsmith, J.: The syllable. In: Goldsmith, J., Riggle, J., Yu, A. (eds.) The handbook of phonological theory, pp. 165–196. Wiley, Malden, MA, 2nd edn. (2011)
16. Guest, E.: A history of English rhythms. W. Pickering, London (1838)
17. Hansen, E., Slyh, R., Anderson, T.: Speaker recognition using phoneme-specific GMMs. In: Proc Odyssey Workshop. pp. 179–184. ISCA (2004)
18. Hsieh, C.H., Wu, C.H., Shen, H.P.: Adaptive decision tree-based phone cluster models for speaker clustering. In: Proc INTERSPEECH. pp. 861–864. ISCA (2008)
19. Kessler, B., Treiman, R.: Syllable structure and the distribution of phonemes in English syllables. J Mem Lang **37**(3), 295–311 (1997)
20. Kozhevnikov, V., Chistovich, L.: Speech: articulation and perception. Translation JPRS 30543, Joint Public Research Service, U.S. Dept. of Commerce (1965)
21. Levinson, S., Rabiner, L., Sondhi, M.: An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst Tech J **62**(4), 1035–1074 (1983)

22. Liberman, A., Ingemann, F., Lisker, L., Delattre, P., Cooper, F.: Minimal rules for synthesizing speech. J Acoust Soc Am **31**(11), 1490–1499 (1959)
23. Martin, T., Wong, E., Baker, B., Mason, M., Sridharan, S.: Pitch and energy trajectory modelling in a syllable length temporal framework for language identification. In: Proc Odyssey Workshop. pp. 289–296. ISCA (2004)
24. Mary, L., Yegnanarayana, B.: Extraction and representation of prosodic features for language and speaker recognition. Speech Commun **50**(10), 782–796 (2008)
25. Mitford, W.: An inquiry into the principles of harmony in language, and of the mechanism of verse, modern and antient. L. Hansard, London, 2nd edn. (1804)
26. Olson, H., Belar, H.: Phonetic typewriter. J Acoust Soc Am **28**(6), 1072–1081 (1956)
27. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an ASR corpus based on public domain audio books. In: Proc ICASSP. pp. 5206–5210. IEEE (2015)
28. Rudnicky, A.: CMUdict 0.7b: https://github.com/Alexir/CMUdict. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (2015)
29. Sadjadi, S., Hansen, J.: Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process Lett **20**(3), 197–200 (2013)
30. Saussure, F.: Cours de linguistique générale. Payot, Lausanne; Paris (1916)
31. Schindler, C., Draxler, C.: Using spectral moments as a speaker specific feature in nasals and fricatives. In: Proc INTERSPEECH. pp. 2793–2796. ISCA (2013)
32. Selkirk, E.: The syllable. In: van der Hulst, H., Smith, N. (eds.) The structure of phonological representations, vol. vol. 2, pp. 337–384. Foris, Dordrecht (1982)
33. Shoup, J.: Phonological aspects of speech recognition. In: Lea, W. (ed.) Trends in speech recognition, pp. 125–138. Prentice-Hall, Englewood Cliffs, NJ (1980)
34. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. Speech Commun **46**(3-4), 455–472 (2005)
35. Siu, M.H., Yu, G., Gish, H.: An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In: Proc ICASSP. vol. 2, pp. 189–192. IEEE (1992)
36. Soldi, G., Bozonnet, S., Alegre, F., Beaugeant, C., Evans, N.: Short-duration speaker modelling with phone adaptive training. In: Proc Odyssey Workshop. pp. 208–215. ISCA (2014)
37. Sugiyama, M., Murakami, J., Watanabe, H.: Speech segmentation and clustering based on speaker features. In: Proc ICASSP. vol. 2, pp. 395–398. IEEE (1993)
38. Wallis, J.: Grammatica linguae Anglicanae. L. Lichfield, Oxford (1674)
39. Wang, G., Wu, X., Zheng, T.: Using phoneme recognition and text-dependent speaker verification to improve speaker segmentation for Chinese speech. In: Proc INTERSPEECH. pp. 1457–1460. ISCA (2010)
40. Wilcox, L., Chen, F., Kimber, D., Balasubramanian, V.: Segmentation of speech using speaker identification. In: Proc ICASSP. vol. 1, pp. 161–164. IEEE (1994)
41. Yamada, M., Pezeshki, A., Azimi-Sadjadi, M.: Relation between kernel CCA and kernel FDA. In: Proc IJCNN. pp. 226–231. IEEE (2005)
42. Yella, S., Motlícek, P., Bourlard, H.: Phoneme background model for information bottleneck based speaker diarization. In: Proc INTERSPEECH. pp. 597–601. ISCA (2014)
43. Zibert, J., Mihelic, F.: Prosodic and phonetic features for speaker clustering in speaker diarization systems. In: Proc INTERSPEECH. pp. 1033–1036. ISCA (2011)