Regroupement par locuteur de messages vocaux

Perrine Delacourt et Christian J. Wellekens
Institut EURECOM, 2229 route des Crêtes,
BP 193, 06904 Sophia Antipolis Cedex, France
{perrine.delacourt,christian.wellekens}@eurecom.fr

Résumé: Cet article traite du regroupement de messages vocaux par locuteur. Dans notre contexte, un message vocal signifie un segment de parole prononcé par un unique locuteur, avec un fond sonore éventuel et d'une durée quelconque. Ayant à disposition une collection de messages émanant de différents locuteurs, il s'agit alors de classifier ceux-ci par locuteurs. La technique de classification mise en œuvre est un regroupement hiérarchique. Elle utilise le rapport de vraisemblance comme critère de regroupement entre deux classes de messages vocaux et le critère d'information Bayesien comme critère d'arrêt global. Les résultats présentés dans cet article montrent l'effacité d'une telle technique.

1 Introduction

La classification par locuteur de messages vocaux n'a été étudiée que très récemment dans la littérature. Cette classification peut être utile à diverses applications. Elle peut servir à sélectionner les messages d'une personne spécifiée laissés sur un répondeur téléphonique ou une boîte vocale. Une fois les messages classifiés, il suffit de choisir le groupe de messages appartenant à cette personne. La classification des messages vocaux par locuteur peut être une étape d'un système de transcription. Le principe est le suivant. Etant donnée une conversation à transcrire (un journal télévisé par exemple), le système de transcription dispose des modèles génériques de parole i.e ces modèles sont entraînés sur un grand ensemble de locuteurs. Pour augmenter le taux de reconnaissance d'un tel système, il est d'usage d'adapter ces modèles à chacun des locuteurs présents dans la conversation. Une des étapes du système de transcription va donc consister, après les avoir isolés, à regrouper tous les segments de parole ou messages vocaux prononcés par un même locuteur (cf [5, 7, 8, 9]). Les données de parole ainsi regroupées servent alors à adapter les modèles de parole au locuteur considéré. La classification par locuteur des messages vocaux peut être également une étape d'un système d'indexation par locuteur d'un document audio (cf [2]). Le document audio est tout d'abord segmenté en locuteurs, puis les segments d'un même locuteur sont regroupés. Si le processus est satisfaisant, l'indexation par locuteur est réalisée. Une amélioration consiste à élaborer des modèles de locuteurs à partir des données obtenues pour chacun des locuteurs.

La présente étude se situe plus particulièrement dans le cadre d'un système d'indexation. Dans ce contexte, nous faisons les hypothèse suivantes : le nombre de locuteurs engagés dans la conversation est inconnu et nous n'avons aucune connaissance a priori sur les locuteurs. En d'autres termes, cela signifie que le nombre final de classes de messages est inconnu et que la technique de classification ne peut utiliser que les données de parole contenues dans les messages vocaux considérés. Pour

répondre à ces hypothèses, nous mettons en œuvre la technique de classification exposée dans [1]. Il s'agit d'un regroupement hiérarchique par agglomération et repose sur l'emploi du rapport de vraisemblance généralisé comme critère de regroupement et du critère d'information Bayésien comme critère d'arrêt du regroupement.

Dans cet article, nous décrivons tout d'abord la technique de regroupement par locuteur de messages vocaux à la section 2. Nous détaillons notamment le critère de regroupement (2.1) et le critère d'arrêt (2.2) utilisés. Nous présentons ensuite les méthodes qui nous permettent d'évaluer les performances de cette technique à la section 3. La section 4 est consacrée aux expériences menées et aux résultats obtenus. Enfin, la section 5 tire les conclusions de cette étude et donne quelques perspectives.

2 Technique de classification par locuteur des messages vocaux

La technique exposée dans cette section vise, à partir d'une collection de messages vocaux (appelés également segments de parole), à regrouper ceux-ci par locuteur. La technique, décrite dans [1], est un regroupement hiérarchique par agglomération (cf [2]). A chaque itération, les deux messages ou classes de messages les plus proches au sens d'un critère de regroupement, sont réunis. Ce processus est réitéré tant que le critère d'arrêt n'est pas atteint.

2.1 Critère de regroupement

Dans notre contexte, le critère de regroupement mesure la probabilité qu'ont deux messages vocaux ou segments de parole d'avoir été prononcés par le même locuteur. Le critère de regroupement que nous utilisons est le rapport de vraisemblance généralisé. Ce rapport a déjà prouvé son efficacité en identification du locuteur [5, 4] ou en segmentation en locuteurs [3]. Etant donné deux segments de parole paramétrisés (deux séquences de vecteurs acoustiques) $\mathcal{X}_1 = \{x_1, ..., x_N\}$ et $\mathcal{Y}_2 = \{y_1, ..., y_M\}$, nous considérons le test d'hypothèses suivant:

- H_0 : les deux segments ont été prononcés par le même locuteur. Leur réunion est modélisée par un unique processus Gaussien : $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y} \sim \mathcal{N}(\mu_\mathcal{Z}, \Sigma_\mathcal{Z})$
- H_1 : chaque segment a été prononcé par un locuteur différent et est modélisé par un processus Gaussien différent : $\mathcal{X} \sim \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})$ et $\mathcal{Y} \sim \mathcal{N}(\mu_\mathcal{Y}, \Sigma_\mathcal{Y})$

Le rapport de vraisemblance généralisé R entre les hypothèses H_0 et H_1 est défini par :

$$R = \frac{L(\mathcal{Z}, \mathcal{N}(\mu_\mathcal{Z}, \Sigma_\mathcal{Z})}{L(\mathcal{X}, \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X}).L(\mathcal{Y}, N(\mu_\mathcal{Y}, \Sigma_\mathcal{Y})} \quad (1)$$

La distance d_R est obtenue en prenant le logarithme de l'expression précédente : $d_\mathrm{R} = -\log R$ (distance est ici un abus de langage car d_R ne vérifie pas les propriétés d'une distance). Une faible valeur de R (i.e. une forte valeur de d_R) indique que l'hypothèse H_1, i.e. la modélisation avec deux Gaussiennes correspond mieux aux données. A l'opposé, une valeur élevée de R (i.e. une faible valeur de d_R) signifie

que la modélisation avec une seule Gaussienne (hypothèse H_0) s'accorde mieux aux données. Dans ce cas, les deux segments ont été prononcés par le même locuteur et peuvent être réunis.

Ce principe est généralisé à des groupes de segments en réalisant le test d'hypothèses entre deux groupes de segments. Nous considérons alors comme séquence de vecteurs acoustiques pour chaque groupe, la concaténation de toutes les séquences de vecteurs acoustiques correspondant à chacun des segments du groupe. A chaque itération, le rapport de vraisemblance est calculé pour chaque couple de groupes de segments. Les deux groupes, dont la valeur de d_R est minimale, sont réunis, à la condition supplémentaire que le critère d'arrêt, décrit ci-après, ne soit pas satisfait.

2.2 Critère d'arrêt

Par hypothèse, le nombre final de classes à obtenir est inconnu. Le critère d'arrêt ne peut donc s'appuyer sur ce paramètre ou sur le nombre de regroupements à effectuer. Le critère d'arrêt a donc un rôle de validation ou d'interdiction de la réunion de deux groupes de segments. Aussi, le critère d'Information Bayesien (BIC) est particulièrement recommandé (cf [1]).

Le BIC est un critère de vraisemblance pénalisé par la complexité du modèle. Avec les mêmes notations que précédemment, le BIC d'un modèle M est déterminé par $\mathrm{BIC}(M) = \log L(\mathcal{X}, M) - \lambda \frac{m}{2} \log N_{\mathcal{X}}$ où $L(\mathcal{X}, M)$ est la vraisemblance de la séquence de vecteurs acoustiques $\mathcal{X}$ pour le modèle M, m est le nombre de paramètres du modèle M et λ le facteur de pénalité. Le premier terme reflète l'ajustement du modèle aux données et le deuxième terme correspond à la complexité du modèle. Ce critère permet de comparer deux modélisations d'un même phénomène. Si nous considérons le même test d'hypothèses que précédemment (cf paragraphe 2.1), le BIC permet de choisir entre la modélisation à une seule Gaussienne (hypothèse H_0) et la modélisation à deux Gaussiennes (hypothèse H_1) pour les deux groupes de segments. La différence de BIC entre ces deux modélisations est calculée comme suit :

$$\Delta\mathrm{BIC} = -R + \lambda P \tag{2}$$

où R désigne le rapport de maximum de vraisemblance explicité à l'équation (1) et le terme de pénalité est donné par : $P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N_{\mathcal{X}}$, d étant la dimension de l'espace acoustique, et λ le facteur de pénalité. Une valeur négative de Δ-BIC signifie que la modélisation avec les deux Gaussiennes correspond mieux aux données. Dans ce cas, la réunion entre les deux groupes de segments considérés n'est pas validée.

En pratique, ce critère est appliqué à chaque itération au couple de groupes de segments susceptibles d'être réunis d'après le critère de regroupement. Si la réunion des deux groupes n'est pas validée, alors l'algorithme de regroupement s'arrête.

3 Méthodes d'évaluation

Pour qu'un regroupement des segments par locuteur soit correct dans le contexte de l'indexation, il doit satisfaire aux conditions suivantes : il doit y avoir autant de groupes de segments N_G que de locuteurs N_L présents dans la conversation et chaque

groupe de segments doit ne contenir que les segments relatifs à un même locuteur et tous les segments de ce locuteur doivent se trouver dans ce groupe de segments. La première condition est facile à évaluer. Il suffit de compter le nombre de groupes de segments obtenus et de comparer ce nombre au nombre de locuteurs effectivement engagés dans la conversation. La deuxième condition peut être évaluée en considérant d'une part le nombre de groupes de segments obtenus et d'autre part la pureté de chaque groupe de segments. Soit $n_{i,j}$ le nombre de segments du groupe i prononcés par le locuteur j et n_i le nombre total de segments contenu dans le groupe i. La pureté p_i du groupe de segments i peut se définir comme suit :

$$p_i = 100 \times \frac{\text{nombre de segments du locuteur majoritaire } k}{\text{nombre total de segments contenus dans le groupe } i}\% = 100 \times \frac{n_{i,k}}{n_i}\% \tag{3}$$

Cette définition est celle donnée par [1]. La pureté p_i nous renseigne sur la proportion qu'occupe le locuteur majoritaire au sein du groupe de segments. D'autres définitions existent et sont commentées dans [2]. Il nous semble peu opportun de calculer cette pureté en termes de nombres de segments. En effet, si la pureté p_i est égale à 95%, nous pourrions en conclure que le groupe de segments est plutôt homogène. Mais si les segments du locuteur majoritaire sont de l'ordre de quelques secondes et l'un ou plusieurs des segments contaminants durent plusieurs minutes, alors il y a de fortes chances pour que le groupe de segments ne soit pas vraiment homogène. Nous proposons donc de remplacer dans la définition précédente, le nombre de segments par le nombre de secondes ou le nombre de trames d'analyse :

$$p_i = 100 \times \frac{\text{nombre de trames du locuteur majoritaire } k}{\text{nombre total de trames contenues dans le groupe } i}\% \tag{4}$$

Nous avons ainsi une idée plus juste de la pureté des groupes de segments. Cette mesure de la pureté nécessite la connaissance de l'indexation de référence.

Dans le contexte de l'indexation par locuteurs, nous privilégions un regroupement avec un nombre de groupes de locuteurs le plus proche proche du nombre réel (idéalement égal sinon supérieur de préférence), avec des groupes homogènes de locuteurs (la pureté de chaque groupe doit être proche de 100%) et la durée moyenne des paroles contenues dans un groupe de locuteur doit être conséquente pour permettre à partir de ces données, une éventuelle modélisation de chaque locuteur.

4 Expériences et Résultats

Nous réalisons deux séries d'expériences pour tester l'algorithme de classification présenté dans cet article.

4.1 Evaluation avec des données de référence

Nous appliquons tout d'abord l'algorithme à des données que nous qualifions de référence. Ces données sont constituées de segments purs de parole : les segments ont été délimités manuellement et ne contiennent les paroles que d'un seul locuteur. Les différents types de données de référence que nous utilisons sont les suivants : 10 conversations créées artificiellement en concaténant des phrases extraites de la base

de données **TIMIT** (parole propre, segments courts de 2 à 4 secondes, anglais, 52 minutes), 10 conversations créées artificiellement en concaténant des phrases extraites de la base de données fournie par France Telecom R&D (**CNET**) (parole propre, segments courts de 1 à 3 secondes, français, 34 minutes), 2 dialogues **DIAL** (impliquant deux personnes) créés artificiellement en concaténant des phrases extraites d'une autre base de données fournie par France Telecom R&D (parole propre, segments longs, supérieurs à 2 secondes, français, 13 minutes) et 2 conversations **CONV** créées artificiellement en concaténant des phrases extraites d'une autre base de données fournie par France Telecom R&D (parole propre, segments longs, supérieurs à 2 secondes, français, 32 minutes).

Pour paramétriser le signal audio, nous utilisons des vecteurs acoustiques formés de 16 coefficients Mel-cepstraux.

Pour chaque expérience, nous présentons 3 graphes de résultats. Le graphe de gauche est consacré au nombre de locuteurs. Il mentionne le **nombre réel de locuteurs** présents dans le document audio et le **nombre de locuteurs effectivement trouvés** suite au regroupement hiérarchique. Ces nombres sont des moyennes sur l'ensemble des documents audio évalués. Le graphe du milieu présente **la pureté** p_i définie à l'équation (4) exprimée en pourcentage. Enfin, le graphe de droite montre la **durée moyenne en secondes** des groupes de locuteurs résultant.

La figure 1 présente les résultats obtenus avec les données de référence. Ces résultats sont une synthèse de multiples expériences (cf chapitre 9 de [2]). Seules les meilleures performances de l'algorithme sont présentées dans cet article pour les différents types de données. Ces performances ne sont pas obtenues avec les mêmes valeurs de paramètres pour tous les types de données.

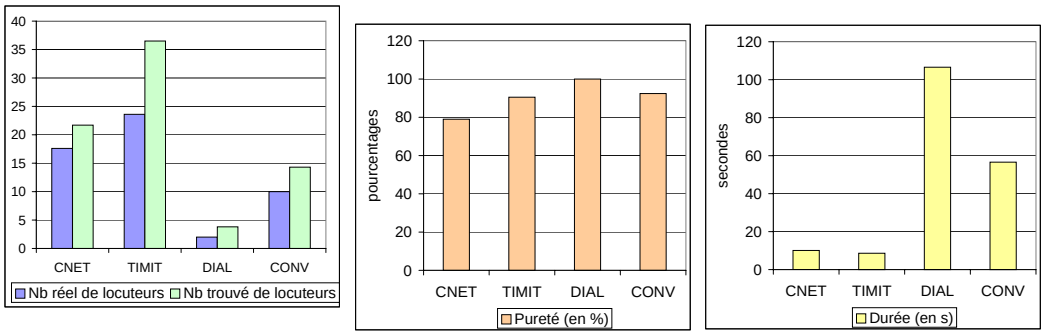

FIG. 1 – *Résultats du regroupement avec des segments de référence pour les données CNET, TIMIT, DIAL et CONV.*

Le paramètre qui importe pour l'algorithme de regroupement est le poids de pénalité λ intervenant dans le Critère d'Information Bayésien (cf équation 2) qui nous sert de critère d'arrêt. Plus la valeur de λ est élevée, plus le regroupement de segments est privilégié. Les résultats expérimentaux montrent que la valeur de λ dépend de la longueur des segments à classifier. Les données CNET et TIMIT sont des documents audio contenant de courts segments (de l'ordre de 1 à 4 secondes). La valeur qui fournit les meilleurs résultats pour cette longueur de segments est 1.0. A l'inverse, les documents audio DIAL et CONV contiennent de plus longs segments

et les meilleurs résultats sont obtenus pour une valeur de λ égale à 1.5.

Dans les résultats que nous présentons (cf figure 1), nous pouvons voir que le nombre de locuteurs obtenu est proche du nombre réel pour les données CNET, DIAL et TIMIT. En ce qui concerne les puretés, celles-ci sont élevées : elles varient de 79% pour les données CNET à 100% pour les données DIAL. Nous constatons des puretés plus fortes pour les documents audio contenant de longs segments (données DIAL et CONV). Ceci s'explique par le fait que la modélisation des segments intervenant dans les critères de regroupement et d'arrêt est plus robuste dans le cas de sgements longs. Enfin, la durée des groupes de locuteurs pour les données DIAL et CONV sont suffisantes pour espérer une modélisation fiable des locuteurs correspondant. Les petites valeurs de durée pour TIMIT et CNET s'expliquent par le fait que le volume de données de parole est faible dans ces documents audio.

4.2 Données issues d'une segmentation préalable

La deuxième série d'expériences concerne des données données issues d'une segmentation préalable. Nous testons l'algorithme de regroupement hiérarchique sur les mêmes données que précédemment : 10 conversations **TIMIT**, 10 conversations **CNET**, 2 dialogues **DIAL** et 2 conversations **CONV**. Ces dernières conversations étant synthétiques, nous testons également le regroupement hiérarchique sur des données réelles : 3 journaux télévisés **JT** français enregistrés dans notre laboratoire (parole préparée et spontanée, 126 minutes) et 49 conversations téléphoniques **SWB** issues de la base de données SWITCHBOARD ([6]) (américain, parole spontanée, durée : de 5 à 10 minutes par conversation). Toutes ces données ont été segmentées en locuteurs à l'aide de l'algorithme **DISTBIC** présenté dans [3]. Aussi, les segments peuvent être impurs, i.e. contenir les paroles de plusieurs locuteurs.

Nous utilisons comme paramétrisation pour le regroupement hiérarchique des vecteurs acoustiques composés de 16 coefficients Mel-cepstraux. Les graphes de résultats sont formés comme précédemment (cf paragraphe 4.1).

La figure 2 compare les résultats du regroupement hiérarchique appliqué aux données de référence et aux données préalablement segmentées. Nous prenons les valeurs des paramètres qui ont fourni les meilleurs résultats pour le regroupement des segments de référence. Pour tous les types de données, nous constatons, comme nous pouvions nous y attendre, une baisse des performances entre la segmentation de référence et la segmentation DISTBIC. Pour les données CNET, le nombre reconnu de locuteurs est en hausse (par rapport au nombre réel des locuteurs) et la pureté et la durée moyenne sont à l'inverse en baisse. Le nombre reconnu de locuteurs étant plus important, la durée moyenne baisse également. La baisse de pureté constatée pour les données TIMIT est liée au fait que les segments résultant de la segmentation DISTBIC sont moins purs que la segmentation de référence. Pour les données DIAL et CONV, la pureté diminue et la durée moyenne augmente entre la segmentation de référence et la segmentation DISTBIC. Par contre, le nombre reconnu de locuteurs baisse dans le cas des données DIAL et augmente dans le cas des données CONV. Pour les données DIAL, la baisse des performances peut s'expliquer par un trop fort regroupement (bien que le nombre de locuteurs reconnus reste supérieur au nombre réel) impliquant ainsi une baisse de la pureté. Mises à part les données CNET,

le nombre de locuteurs obtenus suite au regroupement avec la segmentation de référence et avec la segmentation DISTBIC est quasiment égal pour les autres types de données. Dans les 3 cas, nous pouvons également remarquer que pour les données TIMIT, la baisse de pureté est d'environ 7%, pour les données DIAL d'environ 8% et pour les données CONV de 7%. C'est donc dans les mêmes proportions qu'a lieu la diminution de la pureté pour ces trois types de données.

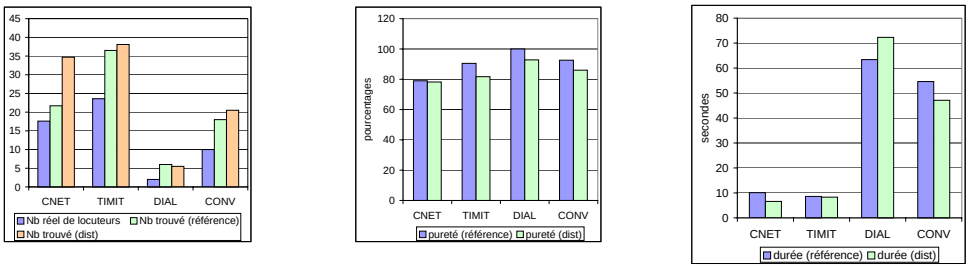

FIG. 2 – *Résultats du regroupement avec des segments issus d'une segmentation préalable pour les données CNET, TIMIT, DIAL et CONV. Comparaison avec les segments de référence.*

Les valeurs optimales du poids de pénalité trouvées pour les segments de référence restent identiques pour le regroupement précédé d'une segmentation en locuteurs. Cette valeur est de 0.8 à 1.0 pour les conversations contenant des segments de courte durée et de 1.5 pour les conversations contenant de plus longs segments.

La figure 3 expose les résultats du regroupement précédé de la segmentation DISTBIC pour les données JT et SWB. Une valeur de 1.2 est prise pour le poids de pénalité λ du regroupement pour les JT et une valeur de 1.5 pour les conversations SWB. Nous constatons par ailleurs une baisse de la pureté pour JT et SWB (respectivement 76.4% et 81.1%) par rapport aux puretés des données synthétiques (cf figure 2). Ceci est essentiellement dû à l'indexation de référence utilisée lors de l'évaluation des performances. En effet, dans le cas de conversations réelles, les segments de locuteurs ne sont pas aussi faciles à délimiter (recouvrement de paroles, traduction simultanée, longs silences intra- et inter-locuteurs, etc...) que dans le cas de conversations synthétiques.

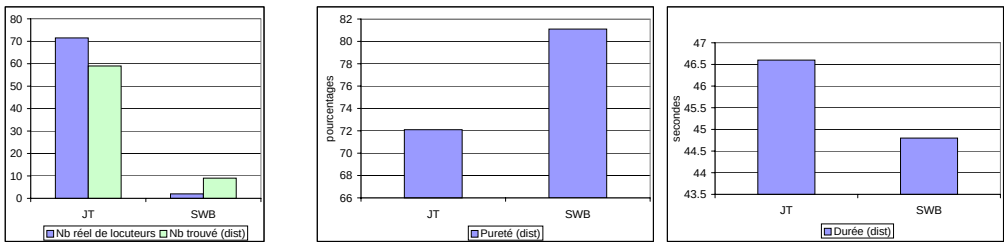

FIG. 3 – *Résultats du regroupement avec les données JT et SWB.*

5 Conclusions et Perspectives

Dans cet article, nous présentons un algorithme de regroupement de messages vocaux par locuteur. Nous mettons en évidence que la valeur du poids de pénalité intervenant dans le critère d'arrêt du regroupement dépend essentiellement de la longueur réelle des messages vocaux ou segments à traiter. L'application de cet algorithme à des données segmentées en locuteurs manuellement d'une part et automatiquement d'autre part, aboutit à une baisse d'environ 8% en termes de pureté. Néanmoins, les bons résultats obtenus permettent une utilisation de cet algorithme au sein de diverses applications. Par ailleurs, l'évaluation de cet algorithme nécessite l'indexation par locuteurs de référence, qui dans certains cas peut poser problème. Nos travaux futurs vont porter plus particulièrement sur le choix automatique de la valeur du poids de pénalité d'une part, et d'autre part, sur une évaluation du regroupement obtenu qui s'affranchirait de l'indexation de référence.

Références

[1] S.S. Chen and P.S. Gopalakrishnan. Clustering via the Bayesian Information Criterion with applications in speech recognition. In *ICASSP*, 1998.

[2] P. Delacourt. *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio.* PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2000.

[3] P. Delacourt and C.J. Wellekens. DISTBIC: a speaker-based segmentation for audio data indexing. *Speech Communication*, 32, Sept. 2000. Special issue on Accessing information in spoken audio, to be published.

[4] H. Gish and N. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, oct. 1994.

[5] H. Gish, M-H. Siu, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 873–876, 1991.

[6] J.J Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 517–520, 1992.

[7] H. Jin, F. Kubala, and R. Schwartz. Automatic speaker clustering. In *DARPA Speech Recognition Workshop*, 1997.

[8] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A. Zissman. Blind clustering of speech utterances based on speaker and language characteristics. In *International Conference on Spoken Language Processing*, volume 7, pages 3193–3196, 1998.

[9] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.