3.中間的統計および平均化オプション

システム出力と参照の比較は、前のセクションで説明したように、固定長間隔またはイベントインスタンスレベルで行うことができます。これにより、パフォーマンスを測定する2つの方法が得られます。セグメントベースのメトリックとイベントベースのメトリックです。それぞれについて、正確な検出を構成するものと、システムがどのようなタイプのエラーを生成するのかを定義する必要があります。これらを、システムの正しい出力と誤った出力を別々に数える中間の統計値と呼びます。問題のポリフォニック特性に基づいて定義します。

3.1。セグメントベースのメトリック

セグメントベースのメトリックは、短時間セグメントでのシステム出力と参照を比較します。各イベント・クラスのアクティブ/非アクティブ状態は、セグメントを表す固定長インターバルで決定されます。アクティビティ表現に基づいて、以下の中間統計が定義されます。

•真陽性：参照およびシステム出力の両方が、そのセグメントでアクティブになるイベントを示します。

•偽陽性：参照は、そのセグメントで非アクティブになるイベントを示しますが、システム出力

それがアクティブであることを示します。

•偽陰性：参照は、そのセグメントでアクティブになるイベントを示しますが、システム出力

それを非アクティブとして示します。

一部のメトリックでは、参照とシステム出力の両方が非アクティブになるイベントを示す場合も、真陰性がカウントされます。真陽性、偽陽性、偽陰性および真陰性の総数は、それぞれTP、FP、FNおよびTNと表される。

時間間隔のサイズは、アプリケーションに必要な解像度に基づいて選択できます。例えば、[31]において、100msのセグメントが提案され、メトリックはフレームベースと呼ばれた。 [26]では、分析に使用される典型的なフレーム長よりも長い時間セグメントのメトリックを計算するために、1秒の長さが示唆された。長いセグメントを評価に使用することで、アクティビティインジケータは長いオーディオチャンクをカバーし、リファレンスとシステム出力の間にある程度のミスアライメントが可能になります。これにより、サウンドイベントの開始およびオフセットのマーキングにおける注釈者の主観性に関連する問題が緩和されます。

3.2。イベントベースのメトリック

イベントベースのメトリックは、イベントごとにシステム出力と対応する参照イベントを比較します。中間統計量は以下のように定義される：

•真陽性：システム出力のイベントで、一時的な位置が、参照の同じラベルを持つイベントの一時的な位置と重複している。襟は、通常、開始およびオフセット、または基準イベント持続時間に対する許容差に対して許容される。

•偽陽性：許可された許容値内の参照内の同じラベルを持つイベントとの対応がないシステム出力内のイベント。

•偽陰性：許容誤差内でシステム出力に同じラベルのイベントとの対応がない、参照内のイベント。

イベントベースのメトリックは、長さに関して実際の時間誤差を測定する場合を除いて、意味のある真のネガティブを有さない。この場合、システム出力および基準の両方がアクティブな事象を含まない時間セグメントの全長が測定される。許容誤差は、所望の解像度に応じて選択することができる。例えば、スピーカーダイアリゼーションでは、データの不正確なラベリングを説明するには±250msのカラーが十分であると推定された。 DCASE2103 [31]では、襟の値は±100msであった。オフセットに許容される許容誤差は、同じ±100 msカラー、または基準イベント持続時間の50％であった。オフセット条件は、非常に短いサウンドイベントと非常に長いサウンドイベントの違いをカバーし、長いサウンドイベントがオフセットが対応するリファレンスイベントから±100ミリ秒以内になくても正しく検出されたとみなすことができます。

3.3。メトリックの計算における平均化オプション

中間的な統計は、クラスに関係なく、グローバルに、または2つのステップでファースト・クラスで、次に個々のクラスの結果を平均して、さまざまな方法で集約できます[33]。高度に不均衡なクラスまたは個々のクラスのパフォーマンスは、2つの平均化方法で計算した場合、全体的なパフォーマンスが非常に異なる場合があります。

インスタンスベースの平均化またはミクロ平均化は、個々の決定に等しい重みを与えます。この場合、各サウンドイベントインスタンス（イベントベースのメトリックの場合）またはセグメントごとのアクティブインスタンス（セグメントベースのメトリックの場合）は、システムパフォーマンスに等しく影響します。の数

Appl。 Sci。 2016年、6月、162/7/17

真陽性（TP）、偽陽性（FP）の数および偽陰性（FN）の数は、試験データ全体にわたって集計され、測定値は全体の値に基づいて計算される。その結果、パフォーマンス値は、考慮する問題の大きなクラスのパフォーマンスの影響をほとんど受けます。

クラスベースの平均化またはマクロ平均化は、各クラスに等しい重みを与えます。個々のクラスに対するシステムのパフォーマンスは、システム全体のパフォーマンスに等しく影響します。この場合、中間データは、テスト・データに対する各イベント・クラスごとに個別に集計されます。集計されたクラス・ワイズ・カウント（TP、FP、FN）は、クラス単位のメトリックを計算するために使用されます。全体的なパフォーマンスは、クラスごとのパフォーマンスの平均として計算されます。これは、考えられる問題のより小さいクラスのシステム動作を強調する値になります。

3.4。相互検証

結果の比較と再現性の重要な側面は、実験的および相互検証のセットアップです。高クラスの不均衡の下での測定バイアスに加えて、マルチクラスの問題では、すべてのクラスがすべての折り返しで表されることを常に保証することはできません[34]。クラスがないと、いくつかのメトリックの計算でゼロ除算が行われます。このような状況を回避するには、フォールトに対する真陽性と偽陽性の合計数を数え、その後に指標[35]を計算することをお勧めします。これはまた、テストされたカテゴリの異なるサブセットにわたる平均化から生じる全体的な平均値の差異（例えば、データを列車/テストフォールドに別々に分割するなど）を回避します。この点で、クロスバリデーションフォールドは、すべてのフォールドをテストした後にのみ最終的なメトリックを計算して、単一の実験として扱う必要があります。これにより、同じデータに対して常にメトリックが計算されます。

4. MetricsforPolyphonicEvaluation

4.1。精度、リコール、Fスコア

Precision（P）とRecall（R）は、情報検索目的で[36]に導入されました

検索の有効性の尺度としてFスコアを導出した。それらは次のように定義されます。

P = TP、R = TP（1）

PrecisionとRecallは情報検索のメトリクスとしては好ましいが、陽性の予測値と感度のそれぞれの分類でも使用されている。これらのメトリックをポリフォニック・データに適合させることは、中間統計の定義に基づいて容易である。 Fスコアは、PおよびRに基づいて計算される：

F = 2・P・R（2）P + R

または、代わりに、中間的な措置に基づいて：

F = 2・TP（3）

2・TP + FP + FN

セグメントベースのP、RおよびFは、セグメントベースの中間統計に基づいて、インスタンスベースの平均化またはクラスベースの平均化を使用して計算されます。計算はパネルAの図4に示されている。イベントベースのP、RおよびFは、イベントベースの中間統計から同じ方法で計算される。計算は図5のパネルAに示されている。同様に定義された事象に基づくP、R、Fは、多音音楽の転写を評価するために使用され、イベントは音符である[21]。

サウンドイベントの検出パフォーマンスを評価するためにFスコアを使用する利点は、広く知られており、理解しやすいことです。その最も重要な欠点は、平均化の選択が特に重要であるということです。 Fスコアの大きさは主に真陽性の数によって決定されるため、インスタンスベースの平均化では大きなクラスは小さなクラスを支配します。クラスベースの平均化では、リコールが定義されていない場合（TP + FN = 0の場合）を避けるために、テストデータにすべてのクラスが存在することを保証する必要があります。現実のレコーディングのデータセットは、不均衡なイベントクラスを持つ可能性が高いため、平均化の選択は常に重要な要素になります。

4.2。誤り率

図4.セグメントベースのメトリクスの計算

図5.イベントベースのメトリックの計算

誤り率は、挿入（I）、削除（D）および置換（S）に関してエラーの量を測定する。セグメントベースのエラー率を計算するために、エラーはセグメントごとにカウントされます。セグメントkにおいて、置換エラーの数S（k）は、正しいイベントが出力されなかった参照事象の数であり、それ以外は何かであった。これは、どの誤ったイベントがどのイベントを代用するかを指定することなく、誤検出と偽陰性の組み合わせによって得られます。 D（k） - 正しく識別されなかった参照イベントの数（置換後の偽陰性が考慮される）およびI（k） - システム出力のイベントの数が正しくない（置換後の偽陽性が考慮される）。これにより、次の式が導かれる。

合計誤り率は、セグメントkの総数に亘ってセグメント毎のカウントを積分することによって計算され、N（k）は、セグメントkの参照においてアクティブとマークされたサウンドイベントの数である。

ER = k = 1 k = 1 k = 1（k）=ΣKS（k）+ΣKD

ΣKN（k）k = 1

（5）

計算結果は図4のパネルBに示されています。MIREX [30]のマルチピッチ転写とDCASE 2013 [20]のサウンドイベント検出の評価には、ポリフォニック検出のエラーレートの同様の計算が使用されました。

イベントベースのエラーレートは、システムイベントの各サウンドイベントの時間的な位置とラベルを参照イベントと比較して、サウンドイベントに関して定義されます。正しい時間的位置であるがクラスラベルが正しくないサウンドイベントは置換としてカウントされ、挿入および削除はシステム出力または参照でそれぞれ正しいかまたは置換されていないサウンドイベントである。全体のエラー率は、エラーカウントに基づいて式（5）に従って計算される。この計算は、図5のパネルBに示されています。誤り率（セグメントまたはイベントに基づく）のクラスワイズ測定は、サウンドイベントクラスごとに個々にエラーカウントを集計するため、置換を考慮することはできませんが、挿入および削除のみをカウントできます。

他の特定の音声認識領域では、誤り率は、評価の目標を反映するために異なって定義される。音声認識では、単語誤り率は単語列に関して定義され、出力には存在するが、参照には存在しない単語として定義され、参照には存在するが出力には存在しない単語として削除され、参照の単語が間違って認識される（出力の中の別の単語）。実際には、1つの置換は1つの挿入と1つの削除を表します（1つの単語が欠けており、もう1つは挿入されています）。 Nは参照内の単語数であるため、単語誤り率は1.0よりも大きくなることがあります。

スピーカーのダイアリゼーションでは、エラーレートは、スピーカーまたはスピーチに正しく帰属しない時間の割合を測定します。時間誤差は以下の通りである：S-話者IDが間違った話者に割り当てられたスコアされた時間のパーセンテージ、I-基準からの非音声セグメントが仮定された話者に割り当てられたスコアされた時間のパーセンテージ、およびD-パーセンテージ基準からのスピーカセグメントが非スピーチに割り当てられたスコアリングされた時間の割合。総誤差は、基準音声の総時間に対する時間誤差の割合として測定される。このエラー率は、シングルラベルシステム出力との比較のために、マルチラベルリファレンスが単一ラベルに投影された後で、[29]のサウンドイベント検出を評価するために使用された。時間的誤差は、スピーカIDの代わりにイベントクラスを用いてスピーカダイアライゼーションのように計算された。

ポリフォニック・サウンド・イベント検出でパフォーマンスを測定するためにトータル・エラー・レートを使用する利点は、音声認識とスピーカー・ダイアリゼーション評価で確立されたメトリックと並行しています。その欠点は、パーセンテージではなくスコアであるため、システムが正しい予測よりも多くのエラーを発生させる場合に1を超えることができることです。これは、能動的な事象を出力しないことによってエラー率1を得ることは自明であることを考慮すると、解釈を困難にする。